Essays on Education Economics

by

Daniela Morar

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2018

Doctoral Committee:

Professor John Bound, Chair
Professor Charles C. Brown
Professor Jason Owen-Smith
Associate Professor Kevin Stange

Daniela Morar

morard@umich.edu

ORCID iD: 0000-0002-5671-8738

Pentru părinții mei, Aurina și Valer Morar.

“Orice român poate fi și campion
Și poate-ajunge
Mai sus de locul doi” (Voltaj)

ACKNOWLEDGEMENTS

First of all, I would like to thank the universe for giving me the opportunity to get this far in life. I feel very fortunate that I got to meet so many wonderful people along the way and work towards making the world a better place. Words cannot express my gratitude towards the following people. I keep you all close to my heart and I will be forever thankful for all your support.

I am extremely grateful to my advisor John Bound for his guidance throughout the program. Charlie Brown, thank you for all the helpful suggestions for my research, and life advice. Kevin Stange, thank you for helping me narrow down my research questions, helping me write better, and keeping my focus on the end goal. Thank you, Jason Owen-Smith for welcoming me into the IRIS family and providing me with the opportunity to start my work on my dissertation. Maggie Levenstein, thank you for all the advice and for providing me with so much guidance!

I would like to thank the IRIS team for their incredible support during my time at Michigan: Nancy, thank you for all the smiles and the candy, Natsuko, thank you for the support with the data, Najla, thank you for the chats, Jinseok, thank you for all the encouragements to not give up on the job market! Sang Teck, thank you for all for teaching me about the world politics an all the encouragement! Patrick, thanks for the coffee breaks and leg days! Ben Koester, thank you for all the help with the data.

This section wouldn't be complete without mentioning my Reed College advisors. David Perkinson, thanks for introducing me to the world of research and sandpiles. Kim Clausing, thank you for being my role model since my Freshman year of college and for being a source of female empowerment!

This Ph.D. journey wouldn't have been possible without the unconditional support of my lovely friends, from everywhere around the world. I would like to thank my good friends from the Michigan Economics department for everything ranging from conversations regarding research, to happy hours and birthday celebrations. Meera, Gaurav, Feiya, Yeliz, Gail, Salma, Enda, Daniel, Chris, Nitya, Paolo, Sreyoshi, Austin, Ben T., Alex, Dyuti, Gretchen, Ben M., Katie, Minjoon, Bhanu, Teju, Jenny, I will always remember you and cherish our times together in Lorch hall! Gaurav and Meera, thank you for listening to all

my job practice talks repeatedly on Skype! Yeliz and Gail, thanks for being such amazing job market buddies!

My friends from outside the Econ department have helped keep a healthy work-life balance: Natalie, Camila, Vaibhav, Clover, Mamta, Brian, Rachel, Tim, Kathryn, and Kelsey. I would also like to thank the Romanian community for helping me stay close to the motherland. Thanks to my Romanian friends for joining me in celebrating Romanian traditions: Cristina, Corina, Christian, Rebeca, Alexandra, Mihai, and Bogdan.

High school friends and puiuți, Andreea, Diana, thank you for keeping me grounded and showing me that true friendship knows no distance! Friends from New Haven, Jose, Duygu, Prashant, and Sadhana, thank you for welcoming me into this new town!

I would like to also thank my family. My college friends, whom I consider my sisters: Marushka, HJ, Vero and Upa have helped me tremendously through these past 13 years. Thank you so much for helping with my transition to the person I am today! No matter where you reside, you will forever have a place in my heart. Last, but not least, I would like to thank my parents, who have sacrificed a lot to provide me with good education. They taught me that everything is possible with enough devotion and hard work. Au și Vali, vă mulțumesc pentru iubirea voastră necondiționată și pentru tot sprijinul pe care mi l-ați acordat. Vă iubesc și mi-e dor de voi!

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

The three essays in this dissertation are focused on the factors that impact persistence in STEM majors. The first chapter uses administrative data from a large public university to investigate whether having a foreign teaching assistant (TA) in a STEM class affects the outcomes of U.S. undergraduate students . This essay considers both subjective outcomes (the median evaluation scores) and objective ones (the students' course outcomes) and concludes that TAs from non-English speaking countries receive between 0.2 and 0.5 points lower median evaluations scores (on a five-point Likert scale) compared to their native-born counterparts, conditional on the course. However, being taught by a foreign TA does not have a significant impact on the students' objective course outcomes, such as grades, STEM major declaration, and STEM graduation. These findings suggest that evaluations of teaching for foreign TAs should be used with caution as they might not be a clear reflection of teaching quality.

The second chapter, from a work with Margaret Levenstein and Jason Owen-Smith, studies the impact of research experience on STEM graduation rates, where research experience is defined as having been employed on a federally funded grant at a large public university. This essay uses a unique dataset where student academic records were matched with longitudinal administrative data on federal research funding at a large public institution. The results find a statistically significant impact of research experience on the probability of graduating in a STEM major and also on the probability of graduating with any major. In addition, this paper begins to disentangle the gender and race related heterogeneous effects of research experience. The results show that undergraduate research employment helps narrow gender and financial gaps in graduation rates (both general and STEM). The findings of this paper indicate potential benefits to students of matriculating in more research-intensive environments and the possibility of interventions to improve the representativeness of STEM population.

The third chapter, from a work with Margaret Levenstein and Jason Owen-Smith, analyzes at the effects of students' socio-demographic and academic characteristics on the necessary and weakly sequential stages to achieve a STEM degree: taking a STEM course in the first year, declaring a STEM major, and graduating with a STEM major. By us-

ing model similar to that of Heckman and Smith (2004), this essay compares the STEM trajectories of male and female students and discusses the effects of different student characteristics on each stage of persistence in STEM. The results show that being a female decreases the likelihood of taking a STEM class in the first year and declaring a STEM major. The largest difference between men and women was in the declaring a major stage, where women were about 9 percentage points less likely to declare a STEM major. This essay also presents decompositions of the effects of gender, race, financial aid status, ACT scores and high school grade point average on each stage leading towards the completion of a STEM degree. These findings suggest that exploring the different mechanisms affecting the differential propensities of male and female students to major in STEM could help reduce the underrepresentation of women in STEM fields.

CHAPTER I

Foreign instructors and student STEM outcomes

Abstract

The past decades have seen an increase in the enrollment of foreign-born students in U.S. STEM (Science, Technology, Engineering, and Math) graduate programs. This paper investigates whether having a foreign teaching assistant (TA) in a STEM class affects the outcomes of U.S. undergraduate students. I consider both subjective outcomes (the median evaluation scores) and objective ones (the students' course outcomes). I use administrative data from a large public university where TAs are conditionally-randomly allocated to classes. I find that TAs from countries where English is not the language of instruction receive between 0.24 and 0.52 points lower median evaluations scores (on a five-point scale) compared to their native-born counterparts, conditional on the course type. However, being taught by a foreign TA does not have a significant impact on the students' objective course outcomes, such as grades, STEM major declaration, and STEM graduation. These findings suggest that evaluations of teaching for foreign TAs should be used with caution as they might not be a clear reflection of teaching quality.

JEL-Classification: I20, I23, J16, J15
Keywords: Higher education, teaching assistants, STEM persistence

1.1 Introduction

Globalization has generated an increase in the number of non US-born graduate students attending American universities (Bound, Turner and Walsh, 2009). Between 1980 and 2015, the number of international graduate students has more than tripled, reaching a record high of $350,000$ students (Zong and Batalova, 2016). On one side, the demand from abroad for a U.S. graduate degree has grown rapidly due to higher college completion rates in countries like China and India (Gaulé and Piacentini, 2013). In addition, because of the high transferability of analytical skills, the demand for a U.S. graduate education has been higher for STEM (Science, Technology, Engineering, and Math) degrees (Bound, Turner and Walsh, 2009). On the supply side, large increases in both federal funding for science and public support for graduate education have provided more opportunities to attend graduate school.

These sizable increases in the number of foreign graduate students have, in turn, caused significant increases in the number of foreign teaching assistants (TAs) in American universities. This study analyzes the impact of the increase of foreign TAs on the educational production function at undergraduate level at the large Midwestern university. [1] The TAs are graduate students who hold office hours, teach smaller sections of the course, and grade assignments and exams. TAs are different from faculty in the university setting. While they are less experienced than the senior staff, they may be able to relate better to the undergraduate students since they share more common experiences, being students themselves at the same university. They constitute an important input in university teaching, making up about 15 percent of the post-secondary instructors in the United States (Bureau of Labor Statistics, 2016).[2]

Given the importance of their contribution to the educational production function, several theories have been invoked to justify why TA characteristics might matter for the undergraduate students. Among these theories is the shifting standards theory of stereotyping (Biernat, Manis and Nelson, 1991) which suggests that peoples' judgments are influenced by relative comparisons among social groups. This theory suggests that lower status groups (e.g. women and minorities) have a harder time demonstrating competence (Foschi, 2000; Basow, Phelan and Capotosto, 2006). In the context of this paper, I assume that undergraduate students compare foreign TAs with native TAs when making decisions

[1] Figure 1.1 shows the trend for STEM versus no-STEM foreign graduate students at this university over a period of 13 years.

[2] The authors use the low cost of hiring a TA as one of the potential reasons why TAs make up for a relatively large percentage of instructors. According to Bureau of Labor Statistics (2016), the median annual wage for post-secondary teachers in the U.S. was $\$75,430$, while the mean wage for TAs was $\$34,240$.

about the effectiveness of teaching. The most common challenges faced by international students have been identified as problems with functionality in the English language and problems with adjusting to the American culture (Andrade, 2006; Trice, 2003) and these two issues appear to be the reasons why undergraduate students might treat foreign TAs differently (Plakans, 1997). Based on these previous studies, I assume that foreign TAs differ from their native counterparts in two important dimensions: familiarity with the U.S. culture and their level of English proficiency. To disentangle the effects of cultural and linguistic differences, I consider two categories of foreign TAs, based on whether or not English is an official or de-facto language in their country of origin.[3] In the absence of any indicators of the foreign TAs' English proficiency and assimilation in the American culture,[4] this categorization is a good alternative to disentangle the two ways in which foreign TAs are different from their American counterparts.

Whith this definition in mind, this study explores the effect of the increase in foreign TAs on the subjective (student evaluations) and also objective (persistence in STEM majors) outcomes of the undergraduate students. I use administrative data from a large public Midwestern institution that contains information on all students (both undergraduate and graduate) and the courses they attended in each semester between Fall 2001 and Winter 2014. This data also contains information on the TAs for each course, which allows me to characterize the TAs based on country of origin, while also controlling for other TA characteristics such as race, gender, and teaching experience.

This study focuses on courses taught by TAs in STEM given the large increase in STEM foreign graduate students. Another reason for focusing on STEM is the small percentage of U.S. undergraduate students who major in STEM fields (Xie and Killewald, 2012; Xie, Fang and Shauman, 2015). In large introductory STEM courses, TA-led sessions are one of the few opportunities for undergraduate students to receive small group instruction, so it is important to examine the impact of the large increase in foreign TAs on undergraduate student outcomes. In addition, large introductory STEM courses offer the ideal setting of conditional random assignment of TAs. More specifically, TAs are assigned to each section based on scheduling constraints, both personal and departmental. Thus, at the time of making their choices, both the TAs and the undergraduate students only have access to information about the time and the day in the week of the section. This makes it almost impossible for the undergraduate students to select a section based on the TA, since they cannot see the name of the TAs when signing up for courses. In addition to this,

[3] One caveat to this explanation is the possibility that TAs from English speaking countries might be closer culturally to the native TAs.

[4] Unfortunately, TOEFL (Test of English as a Foreign Language) scores are not available.

I run balancing tests to show that the characteristics of the undergraduate students are independent of the characteristics of the TA teaching the section, which shows that self-selection into sections of the course is not an issue of concern. This conditional random assignment allows drawing causality conclusions about the foreign TAs.

The first outcome considered is the student evaluations of teaching (SETs), which are used by most colleges and universities in the U.S. to make decisions about their instructors (Murray, 2005). These evaluations provide feedback regarding the quality and effectiveness of the instructors (Svinicki and McKeachie, 2010). In addition to reflecting teaching quality, SETS have also been shown to reflect teaching effectiveness irrelevant factors (Carrell and West, 2010), such as gender, ethnicity and age (Stark and Freishtat, 2014; Andersen and Miller, 1997; Basow, 1995; Cramer and Alexitch, 2000; Worthington, 2002). In this paper, I investigate whether the SETs are related to the country of origin of the TAs. I find that a foreign TA from a non-English speaking country has a median evaluation score of overall quality of teaching between 0.24 and 0.52 points lower than an American TA. Even though foreign TAs from English speaking countries get lower evaluation scores, these results are not statistically distinguishable from both the evaluations of native TAs, as well as the ones for TAs from non-English speaking countries.[5]

The evaluation of foreign born TAs is likely to be dependent on both their teaching performance, as well as on other factors such as cultural differences, social skills, and discipline. To test for this, I examine additional evaluation questions regarding the TA effort exerted, course environment and undergraduate student's self-reported learning from the course. Again, I find that TAs from countries where English is not the official or de-facto language are penalized on criteria regarding effort exerted and learning-inducing class environment. However, undergraduate student self-reported learning is not significantly different in sections led by native TAs than in sections led by non-native TAs. These results are consistent with Watts and Lynch (1989) who suggested that that undergraduate students might blame foreign TAs for their poor course performance.

To assess whether evaluations reflect cultural discontent rather than poor teaching skills, I investigate the effect of non-U.S. born TAs on more objective student outcomes, such as grades, declaring a STEM major and graduating in STEM. The results indicate that foreign TAs have no effect on the grade the undergraduate students get in the course. In addition, I do not find any detectable impact of being assigned to a foreign TA in an introductory STEM course on either the probability of declaring a STEM major or the probability of graduating in STEM.

I also show that the lack of impact of foreign TAs on objective outcomes is not driven

[5]Because of the small sample size, the estimates have a low precision.

by the lack of impact of TAs on the undergraduate student outcomes. To establish this, I employ a value-added model framework to test the importance of teaching assistants as an input for students' academic outcomes. Using a random effects model akin to Carrell and West (2010) and De Vlieger, Jacob and Stange (2017), I find substantial variation in student performance across TAs, both in the contemporary class and also in a subsequent class. These results suggest that, while TAs have substantial impacts on undergraduate student outcomes, foreign TAs are not systematically different from native TAs regarding teaching effectiveness.

My findings have broad policy implications and inform us on how having a foreign TA impacts the outcomes of undergraduate students. One of the policy implications is to be more careful when using teaching evaluations as an indicator of teacher quality. My results are consistent with the shifting standards model that implies that undergraduate students evaluate foreign TAs based on preexisting negative stereotypes about their competence as teachers.

The remainder of the paper proceeds as follows: Section 1.2 reviews the previous literature on TA performance. Section 1.3 reviews the data and presents information about the institutional background of the data. Section 1.4 reviews the empirical setting. Section 1.5 presents the main results of the estimation, Section 1.6 presents extensions of the analysis and Section 1.7 presents concluding remarks.

1.2 Existing literature

Most of the previous papers examining the student evaluations of teaching (SETs) study the connection between the gender of the instructor and their rating of teaching effectiveness. Early findings in this literature show mixed results of instructor gender on SETs (Sidanius and Crane, 1989; Basow and Silberg, 1987; Centra and Gaubatz, 2000; Feldman, 1993). However, the more recent and also more rigorous studies provide consistent evidence of female instructors receiving lower evaluation scores than their male counterparts (Miller and Chamberlin, 2000; Bianchini, Lissoni and M., 2013; Boring, 2017; Boring, Ottoboni and Stark, 2016).[6] Rosen (2017) examines `RateMyProfessors.com` data and finds that female professors receive significantly lower ratings than male professors. In addition, undergraduate students reward the instructors who follow these gender norms (Sprague and Massoni, 2005; Dalmia et al., 2005), and penalize the ones who don't (An-

[6]According to MacNell, Driscoll and Hunt (2015), undergraduate students often have different expectations of their instructors, based on their gender. Thus, they expect male instructors to have more "masculine" attributes, such as professionalism and objectivity, while they expect the female ones to be more "feminine"as in having warmth and accessibility.

dersen and Miller, 1997). While the emphasis of previous literature on student evaluations is on gender, little is known on how country of origin impacts the evaluation scores.

Very few previous studies have addressed the efficacy of teaching assistants (TAs), and even fewer have examined the impact of foreign TAs on student performance. The earlier papers on this topic find mixed results of the effect of foreign TAs on undergraduate students' outcomes. Jacobs and Friedman (1988) examine data from three mathematics courses and one business course at a major Midwestern university and find that foreign TAs are just as effective as native TAs when assessing the final examination scores. They also find no significant differences in the ratings of the foreign TAs compared to native TAs and attribute this finding to the extensive TA screening that foreign TAs are required to undertake at the university.

In another earlier study, Norris (1991), analyzes data from three University of Wisconsin-Madison courses (one survey course and two Economics courses) and finds that sections led by non-native English speakers received higher grades. Contrary to this finding, Watts and Lynch (1989) examine data from Purdue University and conclude that international TAs have a negative impact on post-course standardized test scores.[7] Furthermore, they find no statistically significant relationship between foreign TAs and undergraduate student grades, which could indicate that the native TAs were teaching more to the test than the international TAs. None of these early studies, however, present a setting of random assignment of TAs and they control for very few student and TA characteristics.[8]

The more recent papers in this area have examined only economics courses, with the most prominent being Borjas (2000). In this study, 309 undergraduate students in an intermediate microeconomics course at a large public university are surveyed about their introductory economics courses taken and their experiences with the TAs. The questions from the survey were designed to assess English ability and preparation of foreign born TAs for teaching. The findings show that foreign-born TAs have a negative impact on the undergraduate students' grade. However, foreign born TAs that are better prepared than native TAs do not worsen the achievement of the undergraduate students. Given that the surveys were administrated after the undergraduate students received their grades, these results might be driven by the subjectivity of the answers. For example, as Watts and Lynch (1989) suggest that undergraduate students might blame their bad grades on

[7]The test considered was the revised Test of Understanding College Economics which was designed by the American Economic Association to measure the performance of students in introductory economics courses.

[8]Watts and Lynch (1989) only control for student SAT scores and no additional TA characteristics besides being foreign. Norris (1991) controls for TA experience and high course load, but not any undergraduate student characteristics. Jacobs and Friedman (1988) controls for undergraduate students' SAT scores and the TAs' teaching experience.

foreign TAs, and thus modify their answers to the survey accordingly. Furthermore, the empirical strategy presented in the research does not take account of any additional undergraduate student or TA characteristics.

Following up on this work, Fleisher, Hashimoto and Weinberg (2002) also investigated the influence of foreign-born TAs on undergraduate students and found little adverse effect on the grades in the courses, which the authors argue is a result of the full year of training that the TAs at the university had to undergo. This explanation is consistent with previous studies that show that training leads to both higher ratings from the undergraduate students (Shannon, Twale and Moore, 1998) and a higher sense of self-efficacy[9] towards teaching (Prieto and Altmaier, 1994). Furthermore, Fleisher, Hashimoto and Weinberg (2002) also found that foreign-born TAs got lower ratings in students' evaluations of teaching. One explanation brought forth by the authors is that the international TAs might provide a less desirable class environment due to the cultural gap between themselves and the American-born undergraduates or differences in teaching style.

Additional TA characteristics, besides country of origin, were also found to be relevant for undergraduate student performance measures. Among these characteristics, the most researched one is gender. The studies that analyzed the impact of gender on the instructor on undergraduate student outcomes found mixed results when examining a variety of outcomes, among which grades, persistence outcomes (i.e., dropping the course, taking additional courses in the same field, majoring in that field), and attaining an advanced degree (Robst, Keil and Russo, 1998; Canes and Rosen, 1995; Rask and Bailey, 2002; Bettinger and Long, 2005; Rothstein, 1995; Price, 2010). However, the results on gender matching between TAs and undergraduate students were more indicative of role-model effects: female undergraduate students who have a female TA are less likely to drop out of the course, with no overall effect on performance in the class (Butler and Christensen, 2003). Another strand of the literature found positive impacts of racial/ethnic matching between undergraduate students and instructors (Price, 2010; Lusher, Campbell and Carrell, 2015; Fairlie, Hoffmann and Oreopoulos, 2014).

This study contributes to the literature on student evaluations and foreign TAs by using rich administrative student data from a public Midwestern institution. In comparison with previous studies, I examine a multitude of STEM courses, using a larger sample of undergraduate students. In addition to this, the institutional setting offers a close to random assignment of TAs to course that allows me to draw causal inferences about the impact of these TAs and undergraduate students' outcomes. I also examine a broad range of outcomes of the undergraduate students, which include shorter term ones such as the grades

[9]The term self-efficacy refers to a person's belief in their ability to accomplish a task (Bandura, 1982).

in courses and declaring a major and also longer term ones, such as graduation rates. In addition to the course outcomes of the undergraduate students, this study analyzes outcomes relating to the evaluations of teaching for the TAs, thus bridging the gap between the two existent study areas. Given the large increase in foreign TAs over the past decade and given that this increase is significantly larger in STEM, it is important to analyze the impact of foreign TAs in the context of these large STEM courses that the undergraduate students take.

1.3 Institutional background and data

This section describes the institutional background and data used for my analysis.

1.3.1 Institutional background

I use administrative student data from a public Midwestern institution, where the main colleges are the College of Arts and Sciences (which has approximately 60% enrollment) and the College of Engineering.[10] Teaching at the university is done on a semester calendar system, with Fall and Winter semesters, followed by two shorter Spring/Summer semesters.

In addition to the primary faculty member in charge of leading the main lectures, most large introductory courses also have a TA involved in the instruction of the course. The majority of the TAs are current graduate students enrolled at the university. There are some rare instances where undergraduate students are also allowed to teach, but I only consider graduate students in my analysis. The TAs responsibilities vary based on the course and the department and they involve a combination of grading assignments, guiding discussion or laboratory sections, assisting with the preparation of course materials or leading study sessions. This study only considers introductory STEM courses that the undergraduate students take in their first two terms of classes. I denote as STEM all the fields thought to contribute to technological innovation (Xie, Fang and Shauman, 2015). Although there are various STEM definitions, I employ the one used by U.S. Immigration and Customs Enforcement (ICE) for allowing special work visas for foreign nationals in STEM fields (Gonzalez and Kuenzi, 2012). Unlike the STEM definition used by the National Science Foundation (NSF), the ICE definition[11] doesn't include the social sciences. Given that most social sciences recruit graduate students based on very different criteria than the sciences, I

[10]The College of Engineering has a separate admission process, but the students in this college can take courses from all the other colleges of the university.

[11]https://www.ice.gov/doclib/sevis/pdf/nces_cip_codes_rule_09252008.pdf

believe that using the ICE definition is the better approach. Furthermore, I use the original ICE definition of STEM and disregard additions to the list of STEM degrees in 2011 and 2012 (when fields like psychology, agriculture, etc. were added to the STEM list).[12]

When applying for a TA position in the STEM courses considered, each graduate student specifies their top preferences regarding which courses they would like to teach. These preferences together with the preferences of the faculty of the course are passed on to the person in the department in charge with TA allocations and assignments, which makes the final decision. No screening is involved, but most of the faculty members already know the graduate students (or can ask their adviser about their background). Thus, the faculty members make informed decisions about which TAs would be best suited for their course. The TAs also undergo training prior to their first semester teaching the course or during the first term teaching, depending on the course considered. In addition to this, TAs from undergraduate universities where courses are taught in languages other than English[13] from the College of Arts and Sciences are required to take a college teaching course from the English Language Institute.[14] The departments also provide access to a graduate student mentor, responsible for giving teaching advice and making observations about teaching. TAs are evaluated based on the median score on the teaching evaluations. If a TAs receives a median evaluation score below 3 (on a Likert scale of 1-5) on the question regarding their overall performance (i.e."Overall, the instructor was an excellent teacher."), they receive a warning from the department. If the poor performance is repeated in a subsequent semester, they will no longer be considered for a TA assignment.

Given the various roles TAs can have in teaching, this study considers three possible types of classes: laboratories, discussion sessions and courses which are taught entirely by TAs. Each individual TA has little input in deciding the undergraduate students' grades, and the degree of input the TA has varies slightly by the type of class considered. The laboratories and discussion sessions do not have separate exams, they only have quizzes and laboratory reports, graded solely by the TAs. Since the grade for the course is determined by exams taken in lecture, I match these sections with the grade in the course. Given the large size of these introductory courses, most of the exams are scantron-graded, multiple choice (additional information on the exams is provided in section A.1). In the rare cases of non-multiple choice exams, the TAs get together after the exams and grade together

[12]Section A.1 offers a list of the courses considered, which are introductory courses in: biology, chemistry, physics, mathematics and engineering.

[13]This requirement is waved for students who have received their undergraduate degree from a U.S. based institution or from an institution outside of the U.S. with curriculum in English.

[14]All TAs from non-English-medium undergraduate universities are also required to submit their TOEFL exam scores prior to applying to the respective graduate program.

using an answer key provided by the faculty teaching the lecture. Given these procedures, it is very unlikely that the difference of grades in the sections to be a result of different grading scales. Therefore, using the grade in the course as an outcome should not be viewed as problematic. In addition to this, I consider additional outcomes that would not be influenced by the TAs' ability to influence grades, such as the probability of declaring a STEM major and the likelihood of graduating in STEM.

1.3.2 Data

The data contains all undergraduate students taking classes between Fall 2001 and Winter 2014. The administrative data offers detailed information about the students who are attending this public institution, both undergraduate and graduate students. The data cover the basic demographic information and the entire course taking history of each student. The demographic information includes each student's race (i.e. white, black, Hispanic, Asian, and other (Native American, not indicated, Hawaiian and two or more)), gender (binary male/female), state and country of residency.

For undergraduate students, I use financial aid status in the form of need-based grant eligibility as a proxy for parental income.[15] The largest of the need-based financial grants is the federal Pell Grant, a need-based grant that assists low-income students who are attending universities and other accredited secondary institutions. I create a binary Pell grant variable that identifies students who have received one (or more) of the following grants: Pell Grant, Academic Competitiveness Grant (ACG), Supplemental Educational Opportunity Grant (SEOG) or SMART grant.

I have additional data on Advanced Placement (AP) exams and information about the last high school attended by the student. Since the analysis in this paper focuses on STEM outcomes, I only consider science and math AP tests.[16] In addition to AP test scores, I also control for high school grade point average (GPA), recalculated by the university on a 4.0 scale.[17] SAT and ACT test scores are also included, where SAT scores were standardized into ACT scores using the official ACT conversion table[18].

[15]Data on parental education and income acquired from the admission office has too many missing observations (over 40 percent missing for parental income and over 20 percent for parental education) and multiple imputation methods cannot be used due to the non-randomness of the missing data.

[16]The AP tests considered are: Biology, Chemistry, Physics (Physics B, Physics C: Electricity and Magnetism, and Physics C: Mechanics), Computer Science (Computer Science A and Computer Science AB), Statistics, and Calculus (Calculus AB and Calculus BC).

[17]One caveat is that before 2009, the university included only the courses taken in grades 9-11 for calculating the GPA. After 2009, the university considered all high school courses taken for all grades. However, do not believe that this would be a major issue for my analysis given the richness of my data.

[18]The conversion table can be found at http://www.act.org/aap/concordance/pdf/reference.pdf.

The data also provide information about the courses taken by the undergraduate students each semester. This information contains the course subject and number, the type of course (lecture, discussion session, laboratory, etc.), the number of credits awarded, and the grade obtained in the course. I define a class as a combination of a term (e.g. Fall 2007), course (e.g. Chemistry 101) and lecture.[19] Three dependent variables are used as a measure of students' achievement in a course: the grade in the course, the probability of declaring a STEM major and the probability of graduating in STEM. As explained in Section 1.3.1, the grades considered are the grades in the course taken by the undergraduate student. In the case where the actual section taught by the TA does not have a separate grade, I consider the grade for the course belongs to. I create a binary variable for majoring in a STEM field by using the CIP (Classification of Instructional Programs) codes that identify each major in combination with the STEM definition from the previous section. I use the same method to create a binary variable for graduation with a STEM degree in five years.[20]

I also control for the race and gender of the TA and use the information about each TA's country of permanent residence at the time of submitting their graduate studies application to create a binary foreign TA dummy.[21] I further divide this foreign TA dummy into two categories, based on whether or not they come from a country where English is an official or de-facto language.

In addition to demographic information on TAs, I also have access to data on the student evaluations of teaching (SETs) from Fall 2008 (when online evaluations were introduced) to Winter 2015. For every course the undergraduate students take each semester, they receive an email in the last week of classes with a link to fill out the teaching questionnaires, followed by three reminders. The timing of filling out the evaluations is such that the students evaluate each course before taking the final exam in that course and learning about their grade. Similarly, the TAs do not have access to the teaching questionnaires filled out by the students until the final grades have been released. This "double-blind" procedure insures that TAs do not award grades based on negative evaluations and that the undergraduate students do not rate TAs based on the final exam or their course grade. Furthermore, the evaluations are anonymous and the TAs receive information about their evaluation scores aggregated at section level.[22] Because of the anonymity of the evalua-

[19]For large courses, several lectures might be taught in the same term by different professors. However, TAs are only assigned to one course per term.

[20]Similar results are obtained when considering a six year graduation rate.

[21]In contrast with my study, the U.S. Census Bureau defines a foreign-born person as a person who is not a citizen of the U.S. but resides in the country, or a naturalized U.S. citizen.

[22]The only exception to this are the student comments, which are not aggregated. Unfortunately, I do not have access to these comments.

tions, I cannot identify the individual characteristics of each student submitting the evaluation, but I can identify average demographic information about the students at section level (from the data on the courses the students take).

The teaching evaluation form contains questions regarding the course and all the instructors that taught the course, as shown in Figure A.1. The questions are designated by department (with some being university wide) and type of instructor (primary faculty or TA). Submitting the evaluations is not mandatory and neither is answering every single question on the evaluations.[23] For each question, the student has a choice of five different answers, which the registrar encodes on a Likert scale: Strongly Disagree=1, Disagree=2, Neutral=3, Agree=4, Strongly Agree=5. Given previous research showing that student answers are likely skewed towards either the lower or the higher end, the registrar calculates the median score rather than the mean for each evaluation question and reports it back to the instructors. Section A.4 explains how to calculate the median score for each evaluation question and provides a computational example.

1.3.3 Summary statistics

To estimate the effect of foreign TAs, I consider introductory STEM courses[24] that undergraduate students take in their first two semesters of college. This assures that the undergraduate students have minimal prior knowledge about the TAs and that this is their first exposure to college courses. I restrict the sample to undergraduates who entered as Freshmen and were registered for classes between Fall 2001 and Winter 2014. This is important because I do not include transfer students, whose course taking behavior might be different due to past experience. To study graduation rates, I further restrict the sample to undergraduate students taking classes before Winter 2010 to allow a 5 year graduation rate for the last cohort of undergraduate students that I observe. The sample considered is restricted to American undergraduate students in order to eliminate role-model type of behavior. Furthermore, I restrict the sample to only introductory STEM courses that are necessary to take to declare a STEM major.

The courses are also divided based on the component of the course taught by the TA: discussion session, laboratory or full course.[25] The descriptive statistics are presented in Table 1.1 and they show that laboratories have significantly fewer women than discussion sessions. In general, the sample consists of between 39-48 percent female students, depending on the type of section considered. The sample also consists of almost 70 percent

[23] Even though this practice might introduce selection issues, it is still an important issue to examine.
[24] A complete list of the courses that I select in my analysis is presented in section A.2.
[25] At the university considered, Calculus I and Calculus II are courses taught entirely by the TAs.

white students, about 5 percent black students, 5 percent Hispanics and 15 percent Asian students. The three types of sections that are led by TAs seem to be balanced in terms of race of the undergraduate students and financial aid status. The courses with laboratories have higher average grades and ACT composite scores than the two other type of courses selected. Furthermore, undergraduate students who take courses with labs are more likely to major in STEM and graduate with a STEM major.

Summary statistics for the teaching evaluations sample are presented in Table 1.2. When examining the summary statistics divided by section type, Table 1.2 illustrates that female TAs are less likely to teach a full course than a discussion or lab. The mean age of the TA is approximately 25, the international TAs from English speaking countries make up 5 percent of total TAs in discussion sessions, 8 percent in labs and 16 percent of TAs in full courses. This large variability can be explained by the fact that different departments at the university attract graduate students from various parts of the world (for example, the mathematics department has more students from European countries than the engineering department). About 20 percent of TAs are from non-English speaking countries. The TAs teaching a full course are slightly more likely to have taught more courses before than the other TAs. The median evaluation score for the TA being an excellent instructor is about 4 on a 1-5 scale.

The summary statistics for all TAs (both foreign and native) divided by the country of origin and the type of section is shown in Table 1.3. The first column of Table 1.3 shows that the India is the country with the largest number of foreign TAs from English-speaking countries, while the majority of TAs from non-English speaking countries come from China. A similar pattern is true for laboratories, as shown in the second column of Table 1.3. The analysis for full courses from the third column of Table 1.3 shows that the majority of the international TAs from non-English speaking countries come from China, followed by Japan and South Korea. The majority of foreign TAs from English-speaking countries come from Canada and India. This analysis also shows that the results for TAs from non-English speaking countries might be driven solely by East Asians.

1.3.4 Allocation of TAs into classes

One of the main issues raised when estimating teacher quality is the potential non-random assignment of undergraduate students to courses which would bias the estimates. However, this issue is not relevant to this study. First, there is a conditionally-random assignment of TAs: the undergraduate students choose which section to enroll in, but they only see the name of the TA after courses start. Thus, the undergraduate students only see the time of the day and the day of the week of the section. In addition, the TAs had

no information about the composition of each section before choosing which one to teach. This reduces the potential self-selection of undergraduate students into a section led by a certain TA. Second, my analysis considers only large introductory STEM courses with capped sections. Thus, there is very little room for the undergraduate students to switch among sections or lectures after they learn who their TA will be.

I also use formal tests to analyze the sorting of undergraduate students into classes. A truly random assignment of undergraduate students would imply that all TA characteristics are unrelated to undergraduate student observable and unobservable characteristics. While I cannot directly test for the correlation of TA characteristics with unobservable undergraduate student characteristics, I can explore the sorting of undergraduate students into classes based on observable characteristics. More specifically, I regress the average undergraduate student pre-assignment characteristics on TA characteristics in each section of each course and jointly testing the equality of means (De Vlieger, Jacob and Stange, 2017).[26] I also include term-course-lecture fixed effects (e.g. Fall 2008, Biology 101, Lecture 100) and add time of the class and day of the week of class as controls.

Since the likelihood of having a foreign TA is highly dependent on the STEM field, it is necessary to add course fixed effects in my analysis. One reason for this is that the undergraduate students taking an introductory STEM class in the fall semester might be different than an undergraduate student taking the same class in the winter (or spring) semester, so I also need to account for the semester the course is taken in. In addition to this, I also need to control for undergraduate students taking the same large lecture to make sure that the undergraduate students in the different sections take the same exams and are exposed to the same professor.[27] Furthermore, controlling for the time of the day and day of the week helps remove the possible selection of undergraduate students or TAs who prefer to attend or teach courses early or late during the day or earlier versus later during the week. I cluster the standard errors at the TA level to account for sections being taught by the same instructors over the course of multiple semesters.

Table 1.4 shows the results of these balancing tests. The first panel of the table contains randomness checks for discussion sessions. Columns (1), (2), (4), (5), (6) and (7) show that the being from both an English and a non-English speaking country are not significantly related to the undergraduate students' pre-assignment characteristics, such as gender, race (except for black students), financial status, ACT composite scores. Columns

[26] An equivalent method is performed by regressing each undergraduate student's characteristics on course-section indicators and testing the null hypothesis that the coefficients on the course-section indicators are equal to zero (Braga, Paccagnella and Pellizzari, 2016).

[27] Since there are no large lectures for the courses where the TAs teach the entire course (Calculus I and II), I only control for the course and the term.

(3) and (8) show that TAs from non English speaking countries are marginally less likely to teach black students and students from the state where the university is located. This, however, does not represent a big concern since I control for the undergraduate students' race in all the regressions presented in this study. I also test for differences in assignments of TAs from English speaking countries and TAs from non-English speaking countries and cannot reject the null of no difference (p-values of 0.434 and 0.167, respectively).

I perform similar balance tests for laboratories and full courses, shown in the second and third panels of Table 1.4. For laboratories, black students are marginally less likely to be in discussions lead by non English speaking TAs. When testing for overall differences in assignment to TAs from different countries, I fail to reject the null of no difference (p-value is 0.208). For randomness checks for full courses, English speaking non-American TAs are marginally less likely to teach white students and students with higher ACT composite scores, while non English speaking TAs are marginally less likely to teach Pell grant recipients. When testing for differences in assignments of TAs from English speaking countries and TAs from non-English speaking countries, the only case I fail to reject the null of no differences is for Pell grant recipients. For the remainder of this paper, I control for whether the undergraduate students received a Pell grant in all the regressions presented.

All in all, the balance tables confirm that assignment of TAs into sections is not correlated with observable undergraduate student characteristics, which further informs me that I can credibly estimate the causal effect of the characteristics of the TA on undergraduate student outcomes using least squares regressions.[28]

1.4 Empirical strategy

1.4.1 Course evaluations

In this section, I study the impact of the country of origin of the TA on student teaching evaluations. I estimate the impact of foreign TAs on four important outcomes: the overall TA rating, the degree of effort the undergraduate students believe the TA exerted, the course environment and the self-reported student learning in the course. As explained in the previous section, the question about the overall quality of the TA[29] is the most important question on the TA evaluation questionnaire and it determines the likelihood of the graduate student receiving a teaching assignment in the future. The distribution of the answers for this question is presented in Figure 1.2 and shows that most of the evaluation

[28] I can make this claim by assuming that the student characteristics that are not correlated with observable undergraduate student characteristics are also not correlated with observable TA characteristics.

[29] The question varies slightly across the courses considered, as shown in Table A.1.

scores are between 4 and 5.

I also consider other evaluation categories in my analysis. One important evaluation category is the degree of effort the undergraduate students believe the TA exerted. As seen from Table A.2, these questions relate to how promptly the TA graded assignments, how well they handled questions in the class, how prepared they were for the class, and how knowledgeable they were about the subject taught. In the case that a section contains multiple of these evaluation questions, I take the average of the median answers. Another group of evaluation questions that I consider relates to the course environment (as shown in Table A.3). These questions depend greatly on the course considered, and they relate to how fair the TA was, how willing the TA was to help the undergraduate students outside the class, how enthusiastic the TA was, and whether the TA enjoyed teaching the class. Even though these questions do not relate directly to undergraduate student learning or TA preparedness, I believe they are an important factor in determining the perceptions of undergraduate students regarding the TAs and the country of origin of the TAs. One last category I consider is the self-reported undergraduate student learning in the course. Table A.4 shows the questions from the evaluation form that were selected to indicate how much the students think they learned from the specific course. All these evaluations questions refer to only the section taught by the TA, and not the course as a whole. I use the following regression to analyze the impact of foreign TAs on median student evaluation scores:

$$\begin{aligned} y_{cst} = \alpha_0 + \alpha_1 X_{cst} + \alpha_2 Z_{cst} &+ \gamma_1 \text{Engl speaking foreign TA}_{cst} \\ &+ \gamma_2 \text{Non-Engl speaking foreign TA}_{cst} \\ &+ \rho_{ct} + \epsilon_{cst} \end{aligned} \tag{1.1}$$

I define the outcome y_{cst} to be the outcome for section s, in term t, for course c, which is the median score of teaching evaluation for the four categories considered: the overall quality, the degree of effort the undergraduate students believe the TA exerted, the course environment and the self-reported undergraduate student learning in the course. This score is a section level aggregate score calculated by the institution using the formula for finding the median of a grouped frequency distribution (found in section A.4). The variables of interest are the binary variables indicating a foreign TA from an English speaking county and a foreign TA from a non-English speaking country. The vector X_i contains controls for TA characteristics such as gender, race, age, and Z_{cst} is the vector of controls for the average undergraduate student characteristics in each course c, in section s, in term

t. Since evaluations are anonymous, I can only control for average undergraduate student characteristics in the respective sections of the course. I also include term-course-lecture fixed effects (ρ_{ct}) and add time of the class, and day of the week of class as controls. The standard errors are clustered at TA level.

1.4.2 Undergraduate student course performance

In this section, I present an analogous ordinary least squares model to the one in the previous section, with the scope of analyzing how the TA's country of origin influences undergraduate student outcomes. I employ the following regression model:

$$\begin{aligned} y_{itcs} = \beta_0 + \beta_1 X_i + \beta_2 Z_{tcs} + \gamma_1 &\text{Engl speaking foreign TA}_{tcs} \\ &+ \gamma_2 \text{Non-Engl speaking foreign TA}_{tcs} \\ &+ \rho_{ct} + \epsilon_{itcs}, \end{aligned} \tag{1.2}$$

where y_{itcs} is the outcome measure for undergraduate student i in course c and section s, in semester t. It should be noted that this model is very similar to the model presented in the previous section, with the difference being that I control for individual undergraduate student characteristics, and not section averages like in the previous analysis. The outcomes considered are the grade in the class, ever having declared a STEM major, and graduating with a STEM degree in 5 years. Since the majority of undergraduate students graduate in 5 years as compared 4 years, I allow undergraduate students to take 5 years to graduate. The model considered includes controls for international TAs, both from English speaking countries as well as non-English speaking countries. Once again, the coefficients of interest are γ_1 and γ_2. X_i are the controls for undergraduate student demographics and course taking behavior (gender, race, ACT composite score, high school GPA, financial aid) and Z_{ics} are the controls for TA characteristics such as gender, race, and age.

Given that each undergraduate student could take multiple introductory STEM courses in the first year and given that these courses could be taught by the same instructors (even though not in the same semester), it is necessary to cluster the standard errors at both the undergraduate student level, as well as at the TA level. Cameron, Gelbach and Miller (2011) propose a new variance estimator for OLS that provides cluster-robust inference when there is a two-way clustering that is non-nested. Correia (2016) improves this two and multi-way clustering of standard errors by also allowing for absorption of multiple fixed effects. Therefore, I use the command developed by Correia (2016) to be able to get the correct standard errors for my estimation. Also included in the regression are

term-course-lecture fixed effects (ρ_{ct}).

1.5 Results

1.5.1 Evaluations

The first panel of Table 1.5 provides the estimation results for the overall quality of TAs. The results show that TAs from non-English speaking countries get significantly lower median evaluation scores than native TAs, with a median score between 0.24 and 0.52 points lower. This is a relatively large effect of about half of a standard deviation, with the average across the three samples close to 4.

This effect is only about one third of the effect that Fleisher, Hashimoto and Weinberg (2002) get, but in their research they do not control for other TA characteristics besides country of origin. The estimated effects for non-American TAs from English speaking countries are also negative, although not statistically significant. Interestingly, although female TAs do get lower median evaluation scores than the male TAs in the courses selected, the results are not statistically significant once I control for other TA characteristics.[30] Furthermore, non-white, non-Asian TAs (i.e. blacks, Hispanics, and other races) are also penalized for evaluation scores, with very large effects for the discussion sessions.

Table 1.5 also provides the results of a F-test for the equality of coefficients for the TAs from English-speaking countries and TAs from non-English speaking countries. I fail to reject that the impact of a foreign TA from an English-speaking country on the median evaluation score is the same as the impact of having a foreign TA from a non-English speaking country at a 5 percent significance level.

The rest of the panels in Table 1.5 show the results for the additional evaluation questions considered. The results suggest that foreign TAs from countries that do not have English as their official/de-facto language are perceived as being worse at exerting effort and promoting a desirable class environment. These results are consistent across the different sections considered and significant, except for TA effort in laboratories. These results also show that being a foreign TA from a non-English speaking country lowers the median evaluation score by about half of a standard deviation of the median evaluation scores, where the mean is around 4. Foreign TAs from English speaking countries also get lower evaluation scores as compared with their native counterparts regarding TA effort, but the results are only significant for the courses where they teach the full course.

[30] This result is different from Boring (2017) who finds that students believe that women have a comparative advantage in course preparation and organization of courses, while men have a comparative advantage in class leadership skills.

When estimating the impact of international TAs on the course environment, foreign TAs get lower median evaluation scores and the results are significant for TAs from non-English speaking countries. One last evaluation question that I consider is the one regarding self-reported undergraduate student learning. Except for full courses, none of the results for foreign TAs is statistically significant. This question is also more connected to the results that I present in the next section that involves the undergraduate students' objective outcomes.

Systematically, this set of results show that TAs from non English speaking countries are getting lower median evaluation scores than native TAs on all questions considered except for the ones about undergraduate student learning. The next step is to examine the impact of TA country of origin for both short-term and long-term student objective student outcomes.

1.5.2 Course grade

This section provides the estimation results using the model from the previous subsection. I study the impact of the having a foreign TA on both short-term student outcomes and long-term ones. Table 1.6 shows the estimation results for the ordinary least squares model that has the grade received in the course as the outcome. Each course has letter grades A-E, which are converted to the standard 0-4 scale.[31] I present the results for the three types of TA-led sections that I consider in my analysis. The estimated effect of TAs from non-English speaking countries from Table 1.6 is negative, small and insignificant. The point estimate indicates that having a TA from a non-English speaking country reduces the grade by 0.03-0.04 points, which is one tenth of the difference from a grade to the next one (e.g from B to B+), and it's only around 5 percent of a standard deviation of the grade variable, with a mean of about 3. Besides this effect not being significant, it also constitutes only around one sixth of the effect of one point change in the ACT composite score on the grade in the course. The results indicate that having a TA from a non-English speaking country reduces the grade in the course by 3-4 percent of standard deviation. Even though not directly comparable, these results are slightly lower than the previous results found in the literature, where Lusher, Campbell and Carrell (2015) find that undergraduate students' grades increase between 2 and 4 percent when exposed to TAs of their own ethnicity.

[31] A+, A =4.0 points, A-=3.7, B+ =3.3, B =3.0, B- =2.7, C+ =2.3, C=2.0, C-=1.7, D+=1.3, D=1.0, D-=0.7 and E=0.0

1.5.3 Other outcomes

One concern is that contemporary course grades do not fully capture the full TA effectiveness (Jackson, 2013) and they are just a reflection of different grading policies or standards across TAs. I address this issue by investigating whether having a foreign TAs impacts the undergraduate students' ability for deep learning, a concept used by Carrell and West (2010) to refer to persistent effects of undergraduate student learning. I quantify the effects of deep learning by considering the probability that an undergraduate student ever declared a STEM major and the probability that the undergraduate student graduated with a STEM degree in 5 years. Studying these additional outcomes also addresses any concerns of the TAs having any input on the course grades.

The results using STEM declaration as an outcome are shown in the second panel of Table 1.6. The estimation results show that undergraduate students who have a non English speaking TA in discussion sessions have a slightly higher probability of declaring a STEM major. More specifically, in discussion sessions, having a foreign TA from non-English speaking country increases the probability of majoring in STEM by about 3 percentage points relative to a mean of 60 percent, which corresponds to about 5 percent difference. I am also interested in longer-term outcomes, such as college STEM graduation. The results for five-year graduation rates are shown in the last panel of Table 1.6. The point estimates for country of origin of TA are again very tiny and they indicate no effect of foreign TAs on the undergraduate students' deep learning. All in all, these results indicate that there is no clear evidence that foreign TAs are doing any worse than native TAs in terms of teaching effectiveness, as measured by actual undergraduate student outcomes.

Once again, I also perform F-tests to test whether the impact of having a foreign TA from a non-English speaking country on objective student outcomes is the same as the impact of a TA from an English-speaking country on the same outcomes. For all the three different outcomes considered (grades, probability of declaring a STEM major and probability of graduating with a STEM degree), I find that I cannot reject the equality hypothesis at the 5 percent level.

1.6 Extensions

1.6.1 Robustness checks

I consider the sensitivity of my results to the inclusion of different controls. Table A.11 presents these results. The first column of the table shows the regression results including both the undergraduate student and TA controls, the second column only includes TA

controls, and the last column only includes undergraduate student controls. I present robustness checks only for two of the outcomes considered: median evaluation scores for the overall teaching effectiveness and grades in the course. Across the different specifications considered, we can see that the results are robust to the exclusion of different controls, with the coefficient estimates changing the most when not including TA controls.

1.6.2 Does TA quality matter?

The previous results could be explained by the fact that perhaps TAs don't really affect grades. One method to evaluate the TAs based on their impact on the undergraduate students' grades is the value-added (VA) approach, first implemented by (Hanushek, 1971) and (Murnane, 1975).

The majority of studies relying on the value-added framework have been written in the context of primary and secondary schools (Rockoff, 2004; Rivkin, Hanushek and Kain, 2005; Chetty, Friedman and Rockoff, 2014*a*,*b*; Rothstein, 2010; Hanushek, 1971; Kane and Staiger, 2008). A handful of studies have looked at the variation of professor effectiveness at the university level and found that instructor effectiveness explains a significant share of the variation in undergraduate students' grades (De Vlieger, Jacob and Stange, 2017; Carrell and West, 2010; Brodaty and Gurgand, 2016), subsequent courses (Bettinger and Long, 2010; Figlio, Schapiro and Soter, 2015; Carrell and West, 2010) and labor market outcomes (Braga, Paccagnella and Pellizzari, 2016).

In this section, I provide evidence of the existence of variation in TA effectiveness. I consider the same sample of undergraduate students as in my previous analysis taking two two introductory STEM courses: Calculus I and Calculus II. As explained in section A.2, the exams in Calculus I and II are not multiple choice, but the TAs have very little room for influencing the undergraduate students' grades as the exams are uniform among all sections of the course and the TAs get together to grade (a group of TAs are assigned the same question to grade for all the exams).

Even though value-added modeling (VAM) is an important tool used by researchers, there are conflicting conclusions on the degree of bias and instability of the VAMs (Kane and Staiger, 2008; Rothstein, 2010). One potential factor that could bias the value-added model is the non-random sorting of undergraduate students (Koedel, Mihaly and Rockoff, 2015). Given this concern, balance test were performed (not shown) to assess students' sorting into sections.

I implement my analysis on TA effectiveness in two steps, by using a random effects model similar to the one used by Carrell and West (2010) and De Vlieger, Jacob and Stange (2017). The first step involves estimating the following value-added model using

ordinary least-squares:

$$Y_{ijkt} = \beta_1 X_i + \beta_2 Z_{jkt} + \gamma_t + \theta_k + \epsilon_{ijkt}, \tag{1.3}$$

where I define Y_{ijkt} as the outcome of student i in section j taught by TA k during term t. Here, X_i is the vector of undergraduate student characteristics, Z_{jkt} is the vector of section mean peer characteristics. The regression further controls for unobserved differences in academic achievement across time and grade inflation (γ_t). The coefficient of interest is θ_k, which represents the TA value-added or the contribution of TA k to the performance of the undergraduate students. More specifically, I am interested in the variance of θs across TAs measures the dispersion of TA quality. The corresponding distribution of TA fixed effects is presented in Figure 1.3 and it suggests a large variability in TA effectiveness across the different TAs considered.

The second step is to construct average residuals for each section for each outcome:

$$\tilde{Y}_{jkt} = \sum_{i \in j} \left(Y_{jkt} - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_{jkt} - \hat{\gamma}_t - \hat{\epsilon}_{ijkt} \right) \tag{1.4}$$

The two outcomes I consider are contemporaneous grades (grades in Calculus I) and grades in the follow-up course (the grades in Calculus II of the undergraduate students who took Calculus I). I use the mean residuals to estimate the variance of the TA effects θ_k as random effects with maximum likelihood (using the "mixed" command in STATA with unrestricted covariance matrix).[32]

When modeling the error term in equation 1.3, I assume it is composed of two additive and independent components: a purely random term and a section specific term: $\epsilon_{jkt} = \mu_{jkt} + e_{jkt}$. The section-specific random effects measures common shocks to all undergraduate students in each section, but not common to all classes taught by the same TA. This term is also reflecting the fact that undergraduate students who receive good grades in Calculus I are more likely to receive good grades in Calculus II.[33] Given the two outcomes considered, grade in Calculus I and grade in Calculus II, the error terms can be rewritten as:

[32]Teacher effects are modeled as random effects in Corcoran, Jennings and Beveridge (2011); Konstantopoulos and Chung (2011); Nye, Konstantopoulos and Hedges (2004) and Papay (2011). Random effects models are employed to produce empirical Bayes shrinkage estimators, which are more stable than the unshrunken fixed effects models.

[33]Both De Vlieger, Jacob and Stange (2017) and Carrell and West (2010) assume these common shocks by noting that the estimates of Corr($\theta_k^{\text{Calc I}}, \theta_k^{\text{Calc II}}$) would be biased in the absence of this assumption.

$$\begin{bmatrix} \epsilon_{jkt}^{\text{Calc I}} \\ \epsilon_{jkt}^{\text{Calc II}} \end{bmatrix} = \begin{bmatrix} \mu_{jkt}^{\text{Calc I}} + e_{jkt}^{\text{Calc I}} \\ \mu_{jkt}^{\text{Calc I}} + \mu_{jkt}^{\text{Calc II}} + e_{jkt}^{\text{Calc II}} \end{bmatrix} \tag{1.5}$$

where Calc I and Calc II indicate having taken the respective courses.

Based on this, Equation 1.4 becomes:

$$\begin{bmatrix} \tilde{Y}_{jkt}^{\text{Calc I}} \\ \tilde{Y}_{jkt}^{\text{Calc II}} \end{bmatrix} = \begin{bmatrix} \theta_k^{\text{Calc I}} + \mu_{jkt}^{\text{Calc I}} + e_{jkt}^{\text{Calc I}} \\ \theta_k^{\text{Calc I}} + \theta_k^{\text{Calc I}} + \mu_{jkt}^{\text{Calc I}} + \mu_{jkt}^{\text{Calc II}} + e_{jkt}^{\text{Calc II}} \end{bmatrix} \tag{1.6}$$

The key parameters of interest are the estimates of variances and correlations of Calculus I TA effects for the grades in both Calculus I and Calculus II, which are: SD($\theta_k^{\text{Calc I}}$), SD($\theta_k^{\text{Calc II}}$) and Corr($\theta_k^{\text{Calc I}}, \theta_k^{\text{Calc II}}$). Table 1.7 reports the main estimates of the variances and correlations of Calculus I TA effects for grade outcomes. A one-standard deviation increase in Calculus I TA quality is associated with 0.14 and 0.13 standard deviation increase in undergraduate student course grades in Calculus I and Calculus II, respectively. Converted to course grade points, this is about half of a grade step (going from A- to A). These results are slightly larger than the results of Carrell and West (2010) (who find 0.05 and 0.13 for the variances) and slightly smaller than the results of De Vlieger, Jacob and Stange (2017) (which are 0.30 and 0.20).

Nonetheless, this substantial variation in TA effectiveness both in the current course and also the subsequent course, suggest that TAs do indeed influence undergraduate students' grades and suggest that prior results in this study cannot be explained by the fact that TAs do not make a difference for undergraduate student outcomes, but by the fact that the country of origin of TAs does not make a difference on the undergraduate students' objective outcomes.

1.7 Conclusion

The goal of this paper is to shed light on the effectiveness of foreign TAs in the education production function by examining both subjective and objective student outcomes. I examine the impact of international TAs in large introductory STEM courses, where TAs are conditionally-randomly assigned to sections. This study concludes that foreign TAs are different than native TAs on two important aspects: lacking knowledge of U.S. culture and institutions and worse English language skills. To distinguish between these two effects, I divide the foreign TAs based on the official language spoken in their home country. My study finds that foreign TAs from non-English speaking countries receive systematically lower evaluation scores than native TAs. However, I find no evidence that these differences

translate into differences in grades. Furthermore, when examining longer term outcomes, such as declaring a STEM major and graduating in STEM, I find no evidence that international TAs are detrimental to undergraduate students' measures of deep learning.

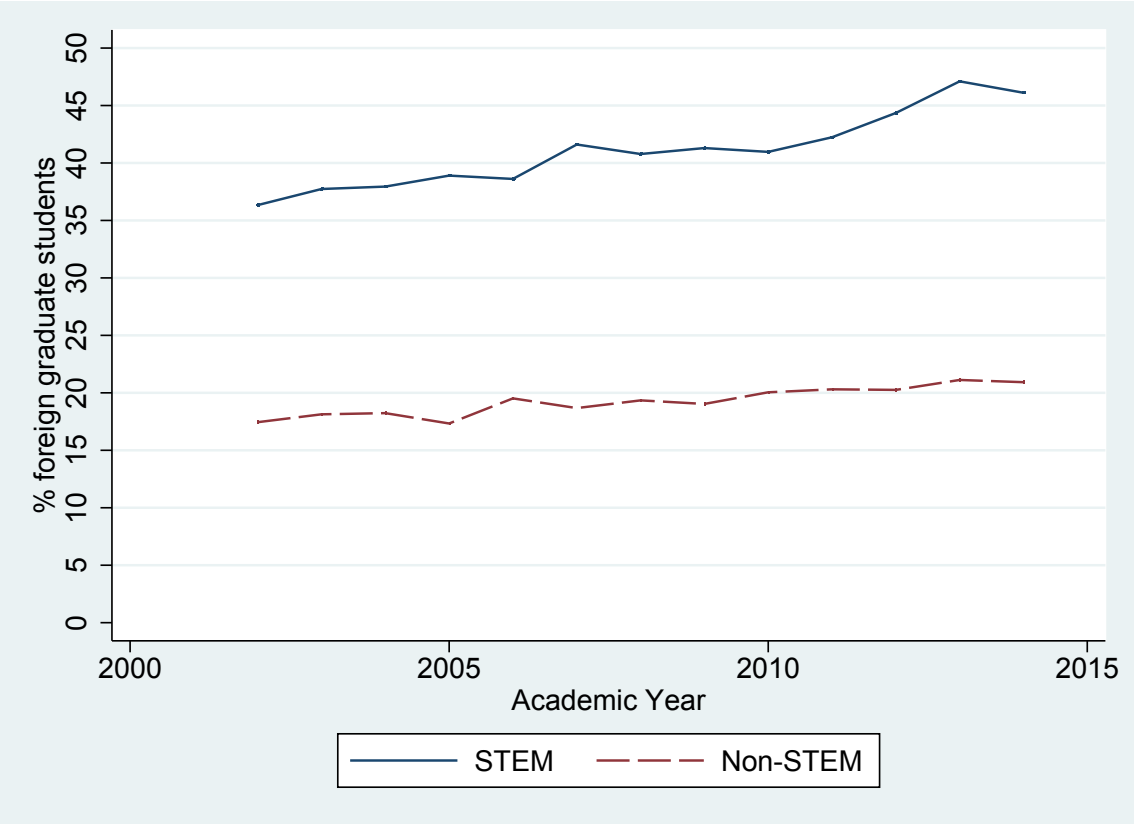
My findings have several implications. First, teaching evaluations should be used with caution as they might not be a clear reflection of teacher quality. These findings support previous findings on student evaluations only being weakly correlated to actual teacher quality (Krautmann and Sander, 1999; Weinberg, Hashimoto and Fleisher, 2009; Carrell and West, 2010; Braga, Paccagnella and Pellizzari, 2014). Second, the fact that foreign TAs receive lower evaluation scores is problematic because it might limit their ability to find an academic job in the future. More research needs to be done on quantifying the actual impact of scores of teaching evaluations on job prospects of international graduate students. In addition to this, international students might be forced to allocate more of their resources towards teaching and away from research so as to increase their evaluation scores.

Another concern, brought up by Mengel, Sauermann and Zölitz (2017) in the context of gender biased evaluations, is the impact of teaching evaluations on the students' confidence. This impact could be driven by stereotype threat, a situation in which the performance of individuals who belong to a negatively stereotyped group is inhibited. Previous literature shows that students with certain immigrant background underachieve in school (Weber, Appel and Kronberger, 2015). In the setting of higher education, the low teaching evaluations scores received by foreign TAs might hinder their ability to teach well in the subsequent semesters. Furthermore, this negative feedback received from undergraduate students might not only affect the foreign TAs' ability and teaching opportunities, but also their interest in an academic job.

All in all, results inform university policy on the existent biases in the student community. In the U.S., as Boring (2017) notes, student evaluations have two main goals: provide feedback on instructional input and help make decisions regarding hiring, firing or promoting instructors. While evaluations could provide some feedback regarding the effectiveness of instructors, the possible existent biases make them unsuitable to be used as "objective" measures of evaluation of instructors.

1.8 Tables and figures

Figure 1.1: Share of foreign graduate students in STEM and non-STEM programs at a large public Midwestern university



Notes: The figure shows the share of foreign graduate students in STEM and non-STEM programs at a large public Midwestern university over 2001-2014.

Table 1.1: Summary statistics for outcomes

	Discussion		Laboratory		Full course	
	Mean	SD	Mean	SD	Mean	SD
Female	0.48	0.50	0.39	0.48	0.40	0.49
White	0.66	0.47	0.68	0.47	0.70	0.46
Black	0.06	0.23	0.04	0.20	0.04	0.20
Hispanic	0.05	0.21	0.04	0.20	0.05	0.22
Asian	0.16	0.36	0.16	0.36	0.13	0.33
Other race	0.04	0.19	0.04	0.19	0.04	0.18
Pell grant	0.20	0.40	0.19	0.39	0.20	0.40
ACT composite score	28.80	3.09	29.22	3.04	28.67	2.74
In state	0.75	0.43	0.73	0.44	0.70	0.46
HS GPA	3.79	0.23	3.82	3.80	3.77	0.24
HS GPA Missing	0.08	0.27	0.07	0.25	0.08	0.28
Grade course	2.80	0.90	3.09	0.82	2.59	0.99
Declared STEM major	0.53	0.50	0.69	0.46	0.51	0.50
Ever graduated with STEM degree	0.42	0.49	0.55	0.50	0.39	0.49
Unique undergraduate students	15256		13957		7729	

Table 1.2: Summary statistics for evaluations

	Discussion		Laboratory		Full course	
	Mean	SD	Mean	SD	Mean	SD
Female TA	0.50	0.50	0.41	0.49	0.25	0.43
White TA	0.62	0.48	0.51	0.50	0.57	0.50
Black TA	0.014	0.12	0.032	0.18	0.013	0.11
Hispanic TA	0.056	0.23	0.074	0.26	0.063	0.24
Asian TA	0.24	0.43	0.32	0.46	0.27	0.45
Other race TA	0.028	0.16	0.047	0.21	0.030	0.17
Age	25.1	2.54	25.7	3.52	24.7	2.13
Foreign TA from English speaking country)	0.056	0.23	0.083	0.28	0.16	0.36
Foreign TA from non-English speaking country	0.18	0.38	0.25	0.44	0.23	0.42
Times taught	4.19	2.37	5.21	2.97	5.76	2.50
Median evaluation score	3.95	0.72	4.05	0.74	4.08	0.75
Number of sections	761		822		300	
Number of unique TAs	191		303		148	

Table 1.3: TAs distribution by country of origin

Countries	Number of TAs		
	Discussion sessions	Laboratories	Full courses
English-speaking countries			
Australia	0	2	3
Canada	2	5	6
Ghana	0	1	0
Hong Kong (China)	0	1	1
India	6	6	6
Israel	0	1	0
Jamaica	0	2	0
Malaysia	0	1	1
Singapore	1	0	2
South Africa	0	1	1
Trinidad & Tobago	1	0	0
United States	142	211	83
Non-English-speaking countries			
Argentina	1	1	0
Brazil	0	1	1
Chile	0	1	1
China	28	51	24
Costa Rica	1	0	0
Colombia	0	1	1
Ecuador	1	0	0
Egypt	0	1	0
Greece	0	1	1
Hungary	1	0	0
Iran	0	2	1
Japan	3	0	0
Mexico	1	0	1
Panama	1	1	0
Peru	0	1	1
Romania	0	0	1
Russia	0	0	2
South Korea	3	6	7
Sri Lanka	0	1	0
Sweden	0	0	1
Taiwan	0	1	2
Thailand	0	1	
Vietnam	0	1	1
Total	191	303	148

Table 1.4: Balancing test of TAs on undergraduate student characteristics for discussion sessions

					Outcomes			
VARIABLES	Avg. female	Avg. white	Avg. black	Avg. Hisp.	Avg. Asian	Avg. Pell	Avg. ACT comp.	Avg. in-state
Discussion sessions								
Foreign TA from non-English speaking country	0.012	0.007	-0.014*	0.000	0.011	-0.017	0.144	-0.029**
	(0.019)	(0.016)	(0.008)	(0.005)	(0.012)	(0.015)	(0.128)	(0.012)
Foreign TA from English speaking country	0.019	-0.033	0.025	-0.000	0.017	0.042	-0.541	-0.023
	(0.022)	(0.039)	(0.027)	(0.008)	(0.019)	(0.046)	(0.412)	(0.025)
F-Test p-value	[0.6786]	[0.8190]	[0.4344]	[0.5736]	[0.9540]	[0.1367]	[0.1397]	[0.1679]
Laboratories								
Foreign TA from non-English speaking country	-0.007	0.001	-0.013*	0.001	0.011	0.016	0.080	-0.019
	(0.012)	(0.018)	(0.007)	(0.006)	(0.012)	(0.014)	(0.098)	(0.013)
Foreign TA from English speaking country	-0.008	-0.008	-0.007	-0.005	0.010	0.013	-0.036	0.002
	(0.017)	(0.026)	(0.006)	(0.008)	(0.016)	(0.016)	(0.120)	(0.021)
F-Test p-value	[0.1706]	[0.9954]	[0.7832]	[0.2064]	[0.5259]	[0.6098]	[0.2447]	[0.2897]
Full courses								
Foreign TA from non-English speaking country	0.001	-0.010	-0.001	-0.008	-0.002	-0.031**	0.182**	-0.013
	(0.019)	(0.022)	(0.006)	(0.006)	(0.014)	(0.016)	(0.083)	(0.016)
Foreign TA from English speaking country	0.018	-0.026*	0.013*	0.005	0.015	-0.002	-0.199**	0.024
	(0.014)	(0.013)	(0.007)	(0.008)	(0.010)	(0.011)	(0.080)	(0.018)
F-Test p-value	[0.4482]	[0.1118]	[0.2081]	[0.5819]	[0.0016]	[0.0589]	[0.3081]	[0.1438]

Notes: Each column is a regression of section-level undergraduate student average characteristics on TA characteristics. We control for foreign TA from a non-English speaking country, as well as foreign TA from English speaking country. Additional controls include the race of the TA, gender, race, teaching experience, and age. All specifications include course-term fixed effects. The robust standard errors are clustered by TA.
The discussion sample size is 761, the laboratories sample size is 822, and the full courses sample size is 300.

Figure 1.2: Distribution of median evaluation scores

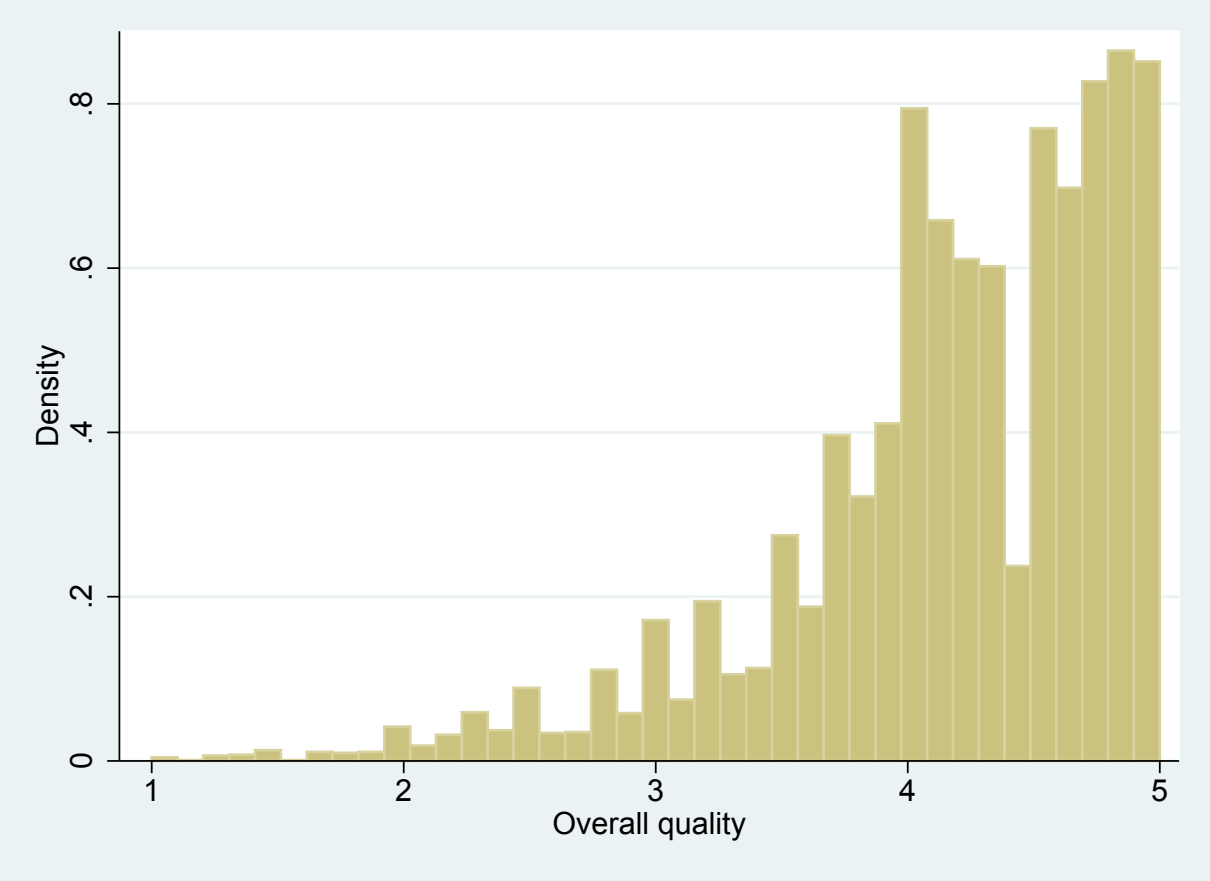


Table 1.5: Results for median evaluation scores (OLS regression models)

VARIABLES	(1) Discussion	(2) Laboratory	(3) Full course
Overall quality of TAs			
Foreign TA from non-English speaking country	-0.36**	-0.24*	-0.52***
	(0.13)	(0.10)	(0.16)
Foreign TA from English speaking country	-0.35	-0.05	-0.27
	(0.19)	(0.21)	(0.15)
F-test for equality of coefficients	0.92	0.34	0.14
Mean dep. var.	3.95	4.05	4.08
SD dep. var.	0.72	0.74	0.76
Observations	763	822	300
TA effort			
Foreign TA from non-English speaking country	-0.29**	-0.17	-0.42***
	(0.10)	(0.09)	(0.11)
Foreign TA from English speaking country	-0.25	-0.05	-0.22*
	(0.13)	(0.18)	(0.11)
F-test for equality of coefficients	0.77	0.50	0.07
Mean dep. var.	3.97	4.102	4.139
SD dep. var.	.59	.57	.50
Observations	761	822	300
Class environment			
Foreign TA from non-English speaking country	-0.29**	-0.24**	-0.29***
	(0.09)	(0.08)	(0.08)
Foreign TA from English speaking country	-0.24	-0.07	-0.16
	(0.14)	(0.16)	(0.09)
F-test for equality of coefficients	0.73	0.24	0.22
Mean dep. var.	4.32	4.20	4.35
SD dep. var.	.44	.55	.37
Observations	761	822	300
Undergraduate student learning			
Foreign TA from non-English speaking country	-0.03	-0.04	-0.22***
	(0.14)	(0.09)	(0.06)
Foreign TA from English speaking country	-0.27	0.01	-0.10
	(0.16)	(0.11)	(0.06)
F-test for equality of coefficients	0.16	0.66	0.07
Mean dep. var.	3.98	3.98	3.94
SD dep. var.	.46	.54	.36
Observations	450	580	300

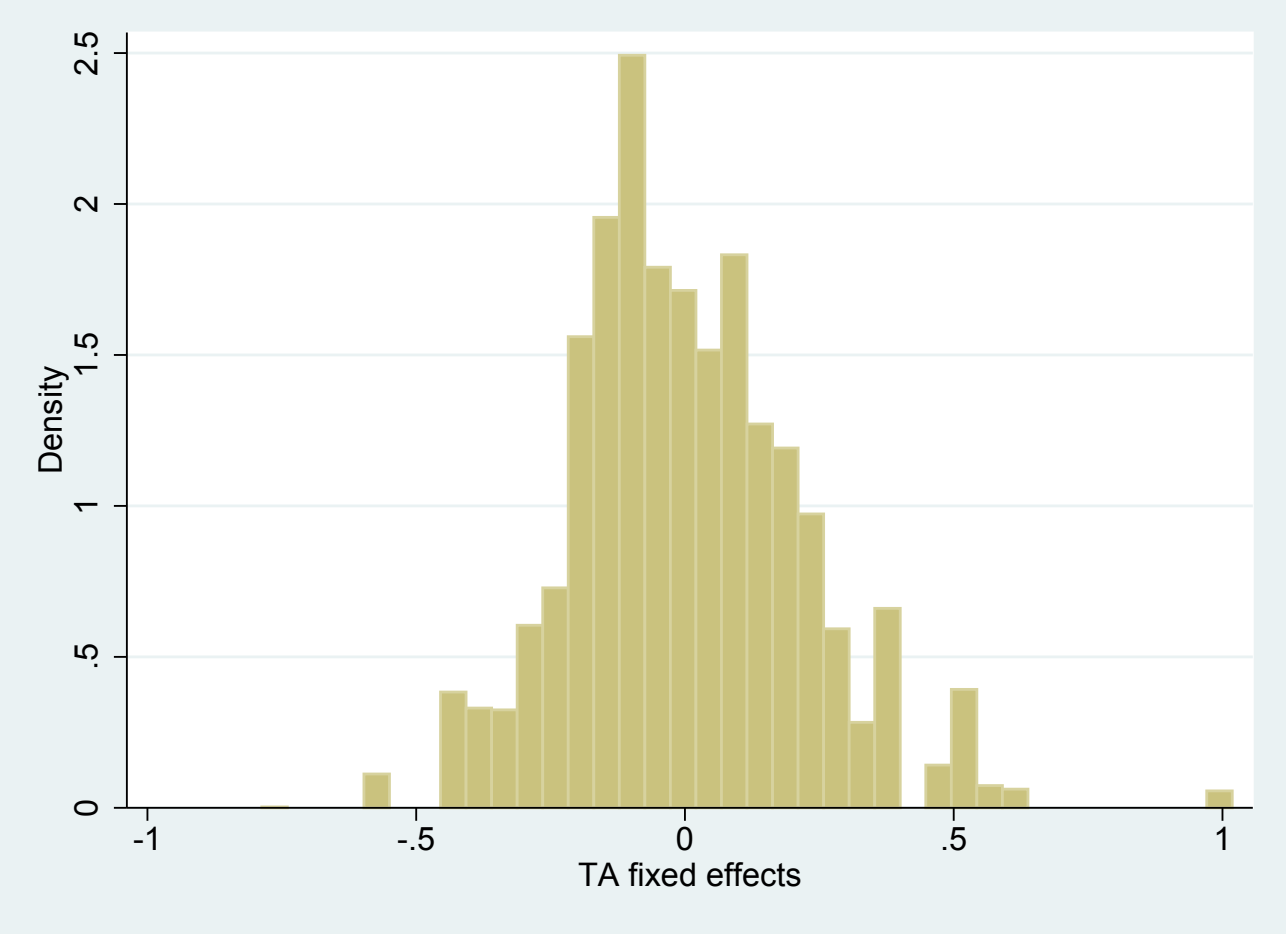
Notes: All specifications control for TA gender, race, age, times taught before, average undergraduate student characteristics, section time, and day of section. Course and term fixed effects are included and the standard errors are clustered by TA.

Table 1.6: Results for undergraduate student outcomes (OLS regression models)

VARIABLES	(1) Discussion	(2) Laboratory	(3) Full course
Grade			
Foreign TA from non-English speaking country	-0.04	-0.03	-0.03
	(0.02)	(0.02)	(0.03)
Foreign TA from English speaking country	-0.05	-0.04	-0.04
	(0.03)	(0.03)	(0.05)
F-test for equality of coefficients	0.69	0.77	0.88
Mean dep. var.	2.80	3.09	2.59
SD dep. var.	0.90	0.82	0.99
Observations	21,800	19,889	8,285
Ever declare STEM major			
Foreign TA from non-English speaking country	0.03*	0.00	-0.01
	(0.01)	(0.01)	(0.01)
Foreign TA from English speaking country	0.00	-0.00	-0.00
	(0.02)	(0.01)	(0.03)
F-test for equality of coefficients	0.09	0.57	0.70
Mean dep. var.	0.53	0.69	0.51
Observations	21,800	19,889	8,285
Ever graduate with STEM degree			
Foreign TA from non-English speaking country	0.00	-0.01	0.01
	(0.01)	(0.01)	(0.01)
Foreign TA from English speaking country	-0.00	-0.01	0.01
	(0.02)	(0.02)	(0.03)
F-test for equality of coefficients	0.70	0.98	0.90
Mean dep. var.	0.42	0.55	0.39
Observations	21,800	19,889	8,285
Unique undergraduate students	15256	13957	7729

Notes: All specifications control for TA gender, race, age, times taught before, undergraduate student characteristics, section time, and day of section. Course and term fixed effects are included and the standard errors are two-way clustered (undergraduate student and TA level).

Figure 1.3: Distribution of TA fixed effects



Notes: The distribution of TA fixed effects is the variance of θ from Equation 1.3.

Table 1.7: Main course grade outcome

TA effect	
SD(Calc I)	0.145
	(0.021)
SD(Calc II)	0.133
	(0.020)
Corr(Calc I, Calc II)	0.758
	(0.185)
Section effect	
SD(Calc I)	0.165
	(0.021)
SD(Calc II)	0.114
	(0.026)
Corr(Calc I, Calc II)	0.420
	(0.224)
Observations	694

Notes: Random effects models are estimated on section-level residuals. First stage models include TA and term fixed effects, in addition to individual controls and section average controls. Residuals are taken with respect to all variables other than TA fixed effects. Robust standard errors clustered by TA in parenthesis.

CHAPTER II

Undergraduate grant employment and persistence in STEM

From a work with Margaret Levenstein and Jason Owen-Smith

Abstract

We study the impact of grant employment on undergraduate graduation rates. We use a unique dataset created by linking course-level student record data with transaction-level data on federal grant expenditures on personnel at a major public university. Our work uses a selection on observables strategy and finds that a paid research experience on a federally funded grant increases graduation rates by 10.1 percent in STEM majors and by 5.5 percent across all fields of study. Undergraduate research employment helps narrow gender and financial gaps in graduation (both general and STEM). Our results indicate potential benefits to students of matriculating in more research-intensive environments and the possibility of interventions to improve the representativeness of STEM population.

JEL-Classification: I20, I23, J16, J15
Keywords: Higher education, teaching assistants, STEM persistence

2.1 Introduction

The United States' strength in terms of productivity, competitiveness and economic growth has been linked to technological development (Xie and Killewald, 2012; Goldin and Katz, 2008; Augustine, 2007). Although a leader in global technology and economy, the U.S. lags behind other developed countries in the number of STEM graduates (Chen, 2013). The National Science Foundation (National Science Board, 2016) finds that only around 9 percent of global university STEM degrees were conferred to U.S. students, while U.S. students hold about 20 percent of the world bachelor's degrees (Ryan and Bauman, 2016; World Bank Group, 2014). There is considerable research that shows that a critical asset to stimulating the country's economic growth is helping students succeed and graduate in STEM disciplines (Ehrenberg, 2010, 2005).

In addition to the low number degrees conferred, STEM fields also suffer from a low participation of women and minorities (Hernandez et al., 2018; US Department of Health and Human Services, 2015; Olson and Riordan, 2012). Despite the fact that women have outnumbered men in college enrollment, there still exists a significant attainment gender gap in STEM degrees (Gayles and Ampaw, 2014; National Science Foundation and Statistics, 2017), with little change since the 1980s (DiPrete and Buchmann, 2013; England et al., 2007; England and Li, 2006; Mann and DiPrete, 2013). According to National Science Foundation and Statistics (2017), women earned about 57 percent of all bachelor's degrees awarded since the late 1990s. More specifically, while women earned more bachelor's degrees in Psychology, biosciences, and social sciences (except for Economics) compared to men, they earned considerably fewer degrees in computer science, engineering and mathematics (National Science Foundation and Statistics, 2017). Participation in STEM also exhibits a racial gap caused by the fact that underrepresented minorities (URM) are less likely than white and Asian students to attain college degrees (Kao and Thompson, 2003). While white and Asian American students are consistently well represented in STEM disciplines (Herrera and Hurtado, 2011; Goyette and Xie, 1999), African-American and Hispanic are underrepresented compared to their overall enrollment (National Science Foundation and Statistics, 2017).

Because interactions between students and faculty members have been linked directly to persistence in college (Terenzini and Pascarella, 1977; Pascarella and Terenzini, 1979; Tinto, 1993), one solution proposed to improve the under-representation of students in STEM fields is participation in faculty-mentored undergraduate research (Hu, Kuh and Gayles, 2007). Working with a faculty member has been shown to improve students' navigation through their STEM major (Cole and Espinoza, 2008; Ullah and Wilson, 2007),

where students are more likely to leave the field due to a "chilly" climate (Seymour and Hewitt, 1997). Furthermore, integration appears to be even more important in the persistence of URM students' (i.e. Black/African American, Latino/a or Native American) in STEM majors (Chang et al., 2014).

There is a large body of literature that focuses on the impact of undergraduate students' employment on their outcomes that find that on-campus employment, as opposed to off-campus employment, has a positive impact on persistence in college (Pascarella and Terenzini, 2005; Hossler et al., 2009). In the federal work-study literature, Scott-Clayton (2011) find that participation in work-study negatively influences the academic outcomes of women (such as first year GPA, first year credits, AA/BA withing four years and dropping out by the fourth year), while it positively influences the outcomes of men. In a follow-up study, Scott-Clayton and Minaya (2016) show that federal work-study also has a positive impact on long-term outcomes, such as bachelor's degree completion and post-college employment, for both female and male students. However, Stinebrickner and Stinebrickner (2003) also study on-campus student employment and find that students who are assigned to jobs requiring more working hours earn lower GPAs.

Studies that examine undergraduate research experience have also found positive effects on student outcomes,[1] such as improved confidence in their science skills (Grandy, 1998; Graham et al., 2013) and a higher knowledge and comprehension of science (Sabatini, 1997) that surpassed the knowledge achieved in ordinary science classes (Ward, Bennett and Bauer, 2003). Furthermore, participating in undergraduate research employment made students more likely to identify themselves as people who "do science", and to improve their oral communication and research skills (Carlone and Johnson, 2007; Hurtado et al., 2009; Barlow and Villarejo, 2004; Bauer and Bennett, 2003; Chang et al., 2014; Seymour et al., 2004; Hunter, Laursen and Seymour, 2007; Kardash, 2000; Russell, Hancock and McCullough, 2007).

Recent studies have also addressed the link between working for a faculty member in college and the students' educational outcomes. These studies show that working with a professor makes students more likely to take advanced courses (Bauer and Bennett, 2003) and graduate at a higher rate (Kim, Rhoades and Woodard Jr, 2003; Gregerman et al., 1998). Students involved in projects with faculty members were also more likely to attend graduate school (Barlow and Villarejo, 2004; Bauer and Bennett, 2003; Pender et al., 2010; Russell, Hancock and McCullough, 2007; Hunter, Laursen and Seymour, 2007; Hathaway, Nagda and Gregerman, 2002). In addition, undergraduate research provided students with a clearer understanding of the type of work involved in a scientific career (Lopatto,

[1] A summary of the literature on undergraduate research experience is offered in Table 2.1.

2010), which increased their desire to pursue STEM-related careers (Bauer and Bennett, 2003; Russell, Hancock and McCullough, 2007; Zydney et al., 2002).

In this study, we evaluate the impact of undergraduate research employment on persistence in STEM, where undergraduate research employment is defined as having been employed on a federally funded grant at a large public university. Despite vast existing empirical studies on undergraduate research, it relies heavily on surveys and fails to support causal claims (Mervis, 2006; Linn et al., 2015; Sadler et al., 2010). In an ideal world, we would like to be able to randomly assign undergraduate employment in a controlled environment to correctly identify and estimate its mean impact on student outcomes. In the context where randomization is not possible, quasi-experimental designs can be used to identify a comparison group that is as similar as possible to the treatment group in terms of pre-treatment characteristics. This method is also used to reduce the selection bias, introduced by the fact that assignment to research experience is also correlated with graduation rates (more motivated or skilled students may be more likely to both do research and graduate) (Rosenbaum and Rubin, 1983; West et al., 2008). We employ a selection on observables strategy, which allows us to construct our counter-factual, the mean outcome of students who worked on a grant had they not worked on a grant. We consider both general and STEM graduation rates. Working with a faculty member gives students a glimpse of the type of work that a scientific career entails (Lopatto, 2010; Kinkead, 2003), which suggests that grant employment should influence STEM graduation rates more than general graduation rates. Our results and consistent with this hypothesis and our preferred matching estimators show a positive and significant impact of approximately 10 percentage points of employment on STEM graduation rates and an impact of approximately 6 percentage points on general graduation rates.

Given that most previous papers have focused on examining small, short-term research programs (Gregerman et al., 1998), we extend the existing literature by examining all paid research positions within a large public university. We also use performance-based evidence (grades, declaring a major, graduating), as opposed to the student self-reported, retrospective accounts of research experience (Hathaway, Nagda and Gregerman, 2002) used by previous papers, which have been shown to be inconsistent with performance-based evidence (Bowman, 2010; Dunning et al., 2003; Feldon et al., 2015). We further contribute to the existing literature by using an innovative dataset that combines administrative student transcript data with longitudinal administrative data on research funding. This unique data on research funding, the UMETRICS dataset, contains information on expenditures made on federal funded grants since 2001. It provides researchers with comprehensive information on employees' salaries, as well as payments to vendors and

subcontractors made from federally funded grants. Thus, linking the UMETRICS data with student transcript data allows us to track all federally funded employment and course taking history for the students who attended this large university over a period of thirteen years (2001-2014). In doing so, we demonstrate some of the value of linked administrative data from universities for understanding both the pathways by which research investments yield returns and the role of high-impact non-classroom experiences in shaping educational outcomes.

Our paper also attempts to disentangle the heterogeneous effects of undergraduate employment based on gender, race and financial status. Given the large differences in STEM persistence rates across students with different socio-demographic backgrounds, it's crucial to test whether undergraduate research employment helps alleviate these gaps. This paper considers two outcomes undergraduate students' outcomes: the likelihood of graduating with a degree in any field and the likelihood of graduation with a degree in a STEM field, conditional on graduation. Our results suggest that undergraduate research employment helps narrow the gender, race and financial gaps in general graduation rates. Furthermore, being employed on a grant helps increase the Black and Hispanic graduation rates, two of the racial groups least likely to graduate from college. We also consider STEM graduation as an outcome and we conclude that grant employment helps decrease the female-male STEM graduation gap, as well as the financial one.

Finally, this study provides estimates of research employment for all types of grant employment, as well as more research intensive positions. In particular, we divide research employment by the amount of research intensity. We define research jobs as the jobs that are related to the student's science career, based on the job description of the university's HR department. Our findings show that research jobs increase STEM graduation rates, and that this effect is higher for female students than it is for male students.

The remainder of this paper is organized as follows. Section 2.2 discusses our data and variable construction. Section 2.3 outlines our identification strategy, Section 2.4 shows our results and Section 2.5 concludes.

2.2 Data

In this section, we describe the data the variable construction.

2.2.1 Institutional context and data

We study the impact of grant employment on graduation rates at a large research university. This large university is home to many colleges, among which are the College of Arts and Sciences (the largest one, making up 60 percent of total undergraduates enrolled) and the College of Engineering.

Our dataset combines information from two different sources: student records data and administrative data on all the federal grants received by the faculty at the university. Both datasets contain a unique individual identifier that allows us to combine the two sources of data to get a complete history of each student's employment and course history.

We use administrative student data from a public Midwestern institution that contains all undergraduate students taking classes between Fall 2001 and Winter 2014. This data contains information on students regarding their demographic characteristics, financial aid status, course outcomes, and degree attainment. The demographic information includes each student's race, gender, and state and country of residency.

The data also provide information about the courses taken by the students in each semester. We have access to the course subject and number, the credit and the grade obtained in the course. We have additional data on Advanced Placement exams and information about the last high school attended by the student, such as grade point average and the name of the high school (identified by its College Entrance Examination Board (CEEB) code). Data on intended major prior to attending the university is also available, as explained in the following section. We define STEM fields as those designated by the U.S. Immigration and Customs Enforcement (ICE), as explained in the following subsection.[2]

The UMETRICS program is a university-specific program that builds upon the federally supported Science and Technology for America's Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness, and Science (STAR METRICS) program.[3] The UMETRICS data are longitudinal administrative data on research funding from 19 major research institutions that contain information on expenditures made on federal awards: payments to individual people, as well as purchases to vendors and sub-contractors. More specifically, the data includes information on each of the grants received by individuals at the university, the number of people employed on each grant, their occupational status and the full-time equivalent (FTE). This comprehensive data contain information on all the employees working on federal grants between 2001 and 2014 at this public institution.[4]

[2]In contrast, the National Science Foundation (NSF) uses a broader definition for STEM fields in which social sciences are also included.

[3]The STAR METRICS project was initiated in 2009 as a partnership between U.S. federal agencies and research universities to measure the impact of federally funded research.

[4]These data are created and maintained by the Institute for Research on Innovation and Science, which

In our data, we cannot identify which positions are work-study positions. [5]

2.2.2 Dataset construction

This section informs the reader on the sample creation and variable construction.

2.2.2.1 Sample creation

We restrict the sample to students who are admitted as Freshmen and remove any transfer students, whose course taking behavior might vary due to past college experience. We only consider students who are enrolled for courses during the time period considered (Fall 2001-Winter 2014). In order to allow students to graduate in 5 years, we restrict the sample to students entering before the Fall of 2010. The resulting dataset of $35,720$ unique students provides a rich source for the analysis of our key questions.

2.2.2.2 Variable construction

Out of the various definitions used by previous literature to define STEM fields, we choose the one designated by the U.S. Immigration and Customs Enforcement agency on April 2008 when the extension for the Optional Practical Training (OPT)[6] was introduced.[7] We do not take into account the additions made to this list in 2011 and 2012 (which include fields like psychology, agriculture, etc.). With this definition in mind, we use the CIP (Classification of Instructional Programs) codes that include all the disciplines offered in academic institutions in the United States to map the majors offered in the university to STEM fields.

These CIP codes are also used to identify students' intent to major in a STEM discipline. We collect information about intended major prior to attending college from three different sources: the Common Application, the SAT, and the ACT. The Common Application asks students to list their areas of interest in college, with no required upper limit for the answers provided. The SAT exam contains a questionnaire on the choice of major, allowing

also makes them available for research use through a virtual data enclave. The full data documentation for the 2017 UMETRICS data release can be found at https://doi.org/10.21987/R7MQ0S. A newer version of the documentation for the 2018 data release can be found at https://doi.org/10.21987/R7GW89.

[5] Most of the standard work-study positions are under the umbrella of the Department of Education, but they do not have a clear code to allow us to identify them.

[6] The Optional Practical Training (OPT) is a period during which undergraduate and graduate students on a student visa are allowed to work for one year.

[7] Information about the OPT can be found at: https://www.ice.gov/doclib/sevis/pdf/nces_cip_codes_rule_09252008.pdf

up to three answers. The ACT asks students to list the college major they plan to have, with only one answer allowed.

Our list of covariates also includes Advanced Placement (AP) tests. Given our interest in STEM outcomes, we only select the science and math AP tests: Biology (BY), Chemistry (CH), Physics (Physics B (PHYSB), Physics C: Electricity and Magnetism (PHYSE), Physics C: Mechanics (PHYSM)), Computer Science (Computer Science A (CSA), Computer Science AB (CSAB)), Statistics (STAT), and Calculus (Calculus AB (CALAB), Calculus BC (CALBC)). In addition to AP test scores, we also have access to high school grade point average (GPA). The university considered recalculated all the high school GPAs on a 4.0 scale. One caveat is that before 2009, the university included only the courses taken in grades 9-11 for calculating the GPA. After 2009, the university considered all high school courses taken for all grades. However, we do not believe that this would be a major issue for our analysis. We also use composite ACT scores as a covariate, converting the SAT scores of the students who did not take the ACT into ACT scores using the concordance tables provided by the College Board.[8]

Another important factor for college persistence is parental education and income (Hellerstein and Morrill, 2011). Unfortunately, data on parental education and income acquired from the admission office contains a very large number of missing observations (over 40 percent for parental income and over 20 percent for parental education) and we cannot use multiple imputation methods due to the non-randomness of the missing data. Instead, we use need-based grant eligibility as a proxy for parental income. Need-based grants have been showed to be good indicators of both the probability that students enroll in college (Deming and Dynarski, 2009), as well as persistence in college (Deming and Dynarski, 2009; Bettinger et al., 2009). The largest of the need-based financial grants is the federal Pell Grant, a need-based grant that assists low-income students who are attending universities and other accredited secondary institutions. We create a binary Pell grant variable that identifies students who have received one (or more) of the following grants: Pell grant, Academic Competitiveness Grant (ACG), Supplemental Educational Opportunity Grant (SEOG) or SMART grant.

Student demographic characteristics also play an important role in their educational attainment. We have information on each student's gender (binary male/female), race (white, black, Hispanic, Asian, and other: native American, not indicated, Hawaiian and two or more) and country/county of residency. We use the information about each student's residence at the time of submitting their college application to define both in-

[8]The concordance tables can be found at: https://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-2009-40-act-sat-concordance-tables.pdf

state/out of state status, as well as international student status (for students with their country of residency outside of the United States).

For comparison purposes, we consider two measures of grant employment. The first measure of research experience considers all the grants from all federal sources listed in the UMETRICS dataset. One problem with this measure is that it includes administrative jobs that might not contribute much to the development of STEM-specific skills. Thus, we also consider an alternative measure of grant employment that includes only jobs that are related to the student's science career, denoted "research jobs". We select these jobs based on the job description from the university's HR department. For example, we categorize positions such as Research Associate, Assistant in Research, and Laboratory Assistant as research jobs, and positions such as Clerk, Library Assistant, and Secretary as non-research jobs.

2.3 Methodology

2.3.1 Treatment and outcome

We estimate the impact of research experience on persistence at a large public institution. Our outcome measure, graduation rate, is calculated based on a five-year window, starting with the first semester of courses attended at the university as a Freshman. We consider both general graduation and STEM graduation rates.

Our treatment variable is undergraduate research experience, a binary variable that measures having been employed on a federally funded grant while attending courses at the university.[9] As seen in Table 2.1, the students at this university are slightly more likely to be working in their later years, but the majority (70 percent) start working in their first year of college.[10] Since we are concerned that students who work on a grant in their senior year are also more likely to graduate, we remove all instances of employment that happen in the student's senior year.[11]

When estimating the impact of grant employment, one must account for a two-sided selection process. First, the student decides whether he or she wants to gain research experience and applies for a job. In the next step, each professor selects one or more applicants to be hired from the pool of all applicants. However, because of the richness of our data and the possibility to control for pre-college interest in STEM (AP exams, high

[9]We exclude all grant employment that takes place before the first semester and after the last semester enrolled for courses.

[10]The density of the number of months employed is shown in Table 2.2.

[11]The Sensitivity Analysis section talks more about this bias.

school GPA, intent to major in STEM) selection on observables is a plausible assumption.

The basic regression we would want to estimate is one where we regress our two outcomes on the treatment status. In an ideal world, we would have a random assignment of treatment and we would not have to add any covariates to the equation. Absent such a set-up, we use matching techniques (Rosenbaum and Rubin, 1983) to compare treatment and control groups in the presence of selection.

Table 2.2 presents the descriptive statistics of the variables used in our analysis for the full sample, as well as for the treatment and control groups. Based on this table, the full sample contains about 52 percent women, with the sample employed on a grant having slightly more women. Our full sample contains about 6 percent black students, 5 percent Hispanic students and 11 percent Asian students. 64 percent of the population comes from the state where the university is located, 19 percent are ever recipients of Pell grants and the average ACT composite score is 28.1. Students who are employed on a grant are more likely to have on average higher high school GPA and ACT composite score and are more likely to have received a Pell grant. Furthermore, the AP scores for science and Math are higher for the students with job experience than for those without it.

The last two columns of Table 2.2 show that more female students who are employed have a research position, as compared to male students. Furthermore, while Hispanics and blacks working on a grant are less likely to have a research job, Asian students are more likely. On average, students who hold research positions are also more likely to graduate. This suggests a positive selection on observable characteristics for employment.

Using this data, we adopt a "selection on observables strategy" (a term adopted from Heckman and Robb (1985)) to estimate the average treatment effect on the treated or the ATET. Formally, we denote Y_1 as the potential outcome for the students employed on a grant, Y_0 as the potential outcome for the students not employed on a grant, and T as the treatment (i.e. grant employment). The observed outcome is $Y = TY_1 + (1-T)Y_0$ and we want to estimate:

$$\Delta_{TT} = E(Y_1 - Y_0|T = 1) = E(Y_1|T = 1) - E(Y_0|T = 1), \tag{2.1}$$

where $E(Y_1|T = 1)$ is the expected grant employment outcome conditional on grant employment and $E(Y_0|T = 1)$ is the expected non-grant employment outcome conditional on grant employment. Identifying $E(Y_0|T = 1)$ is challenging since it is unobservable: we cannot observe the outcomes for the students who did not work on a grant in a world where they had.

2.3.2 Identification

Our matching technique requires the three main conditions. The first one, the conditional independence assumption (CIA) states that once we control for all observable variables, the potential outcomes are independent of treatment assignment (or that $(Y_0, Y_1) \perp T|X$). Rosenbaum and Rubin (1983) show that estimation doesn't require the CIA, but a weaker assumption called the conditional mean independence (CMI), also called the "balancing property". The CMI assumption implies that once we control for the covariates X, the treatment does not affect the conditional mean of each potential outcome:

$$E(Y_0|X, T = 1) = E(Y_0|X, T = 0) = E(Y_0|X) \tag{2.2}$$

Another assumption we need is the common support assumption, namely that for each value of X, there is a positive probability of participation given X, which translates to $0 < P(X) < 1$ for all X. We call this probability the propensity score: $P(X) = Pr(T = 1|X)$.

With this definition of the propensity score, Rosenbaum and Rubin (1983) show that if the CIA assumption holds for X, it also holds for $P(X)$ so that $Y_0 \perp T|P(X)$, and thus:

$$E(Y_0|P(X), T = 1) = E(Y_0|P(X), T = 0) = E(Y_0|P(X)) \tag{2.3}$$

The last assumption is the "Stable Unit Treatment Value Assumption" (SUTVA), which requires that the students who are not employed on a grant are not affected by the treatment. This assumption fails if there are general equilibrium effects generated by spillovers. While we cannot test for the existence of spillover effects, we believe that they are negligible, especially as the treated students are only 10 percent of the total. This assumption also fails if the students are employed on grants from outside of the university considered. We are not too concerned about this possibility since this would bias our estimates downwards, towards a smaller impact of grant employment on persistence in STEM.

2.3.3 Estimation

Given these assumptions, there are many appropriate estimators we could use to calculate the ATET. We begin by considering two different estimators, with pros and cons of using each one of them. The first one, the inverse probability weighting (IPW) estimator, uses weighted averages of the observed outcome variable to estimate means of the potential outcomes.[12] In the estimation process, each weight is the inverse of the esti-

[12]We estimate the treatment effects using the STATA command teffects. This command, unlike psmatch2, calculates the standard errors based on Abadie and Imbens (2012) and takes into account that the propensity score is estimated prior to the matching step.

mated probability that an individual receives a treatment level and it is calculated using the following formula:

$$\hat{\Delta}_{TT} = \frac{1}{N^T}\sum_{i=1}^{N} Y_i T_i - \frac{1}{N^U}\sum_{i=1}^{N}\left(\frac{1}{N^U}\sum_{i=1}^{N}\frac{\hat{P}(X)(1-T_i)}{\hat{P}(X)}\right)^{-1}, \tag{2.4}$$

where N^T is the number of treated units and N^U is the number of untreated units.

Huber, Lechner and Wunsch (2013) and Busso, DiNardo and McCrary (2014) show that the IPW estimator has very good finite-sample properties when compared to other estimators and it requires no assumptions about the functional form of the outcome model. However, IPW can be problematic since it is very sensitive to extreme values of the propensity score and also to small misspecifications.

Therefore, we employ a second estimator, which is the nearest neighbor with replacement. Monte Carlo simulations (Frölich, 2004; Huber, Lechner and Wunsch, 2013; Busso, DiNardo and McCrary, 2014) show that the nearest neighbor with replacement estimator performs very poorly in comparison with other estimators in terms of mean squared error, due to the very high variance of the estimator. The high variance of the estimator is caused by the estimator ignoring all the observations close to the treated units, but not the closest ones. Despite this problem, the estimator exhibits a low bias and it is not sensitive to extreme values of the propensity score, which makes it a good alternative estimator. By increasing the number of neighbors, we increase bias (since we use matches that are farther away), but decrease variance (since we use more untreated units as points of comparison). In our estimation, we use the nearest neighbor with replacement estimator with one neighbor.

2.3.4 Propensity score specification

We first investigate the selection of students based on observable characteristics to explore the common support assumption. In addition to this, this procedure is informative for our sensitivity analysis. We consider different sets of conditioning variables that determine participation in research experience. Covariate selection is a relatively complicated task and there are benefits and costs to increasing the number of covariates. Not including all the important covariates can increase the bias of the estimates, as shown by Heckman, Ichimura and Todd (1997). However, including too many covariates can also be problematic. First of all, we only include the conditioning variables that affect both the treatment and the outcome and exclude the variables that are affected by the treatment. Bryson, Dorsett and Purdon (2002) and Augurzky and Schmidt (2001) find that including

too many covariates that are not significant can increase the variance, even though they will not bias the estimates and it will not make them inconsistent.[13] Moreover, they show that including extraneous variables could lead to a violation of the common support condition.[14] We follow Stuart (2010) and include all the variables previously discussed in order to reduce the bias causes by not including the relevant variables.

We use institutional knowledge, previous empirical findings, and economic theory to determine the best choice of covariates. Previous research focused mainly on student characteristics and experiences (both from high-school and college) to disentangle the factors that influence persistence in STEM. Since high school class rank has been shown to be among the most important determinants of college success in STEM (Ellington, 2006), we use high school GPA quantiles[15] in our analysis. Furthermore, we also include ACT scores,[16] and a vector of different AP tests with scores of at least 3. Because interest in math and science are strong indicators of persistence in STEM (Maple and Stage, 1991; Mau, 2003; Tai et al., 2006; Maltese and Tai, 2010, 2011), we only use AP science and math test scores. Another pre-college variable that we include is a binary variable for interest in a STEM major. By including both AP tests and intent to major in STEM, we control for possible selection into STEM research jobs.

We also include an indicator for Pell grant receipt (and Pell grant gender interaction) to account for the large differences in STEM persistence among students from different SES backgrounds (Schneider, Swanson and Riegle-Crumb, 1998; Miller and Kimmel, 2012). The propensity score specification further includes demographic characteristics (gender, race, gender-race interactions), in-state status (as measured by the address of the student at the time of enrollment), international student status and cohort fixed effects. The estimates of the propensity score model with a probit functional form for both treatments considered are shown in Tables 2.3-2.4.

The first column of each table shows estimates from a logistic regression, while the second column presents marginal effects estimated at the mean of observable characteristics. The regressions show that gender is a statistically significant predictor of research experience. Being female, Asian, and a Pell grant recipient significantly increase the likelihood of participating in research experience. In addition, students with higher high schools GPAs, ACT composite scores, and students from the state where the university is located are also

[13]This is true in any multivariate setting, not just propensity score estimation.

[14]One test suggested for solving the selection of covariates for the propensity score problem is to start with a simple model and keep adding variables. Then the variables are kept if they are statistically significant and if they increase the prediction rate (Black, Daniel and Smith, 2005).

[15]The university converts all the high school GPAs received on a 4.0 scale.

[16]We consider ACT composite scores (we convert SAT scores as explained above).

more likely to be employed on a grant. Having taken an AP science or math test is also a very strong predictor of research experience. More precisely, students who have taken the AP tests in Biology (BY), Chemistry (CH), Physics B (PHYSB),[17] Physics C: Mechanics (PHYSM) and Calculus AB (CALAB)[18] are more likely to be employed on a federally funded grant. Interestingly, black and other race students are more likely to have been employed on a grant, but not more likely to hold a research-intensive position.

2.3.5 Balancing tests

We perform balancing tests to check that, at each value of the propensity score, the covariates chosen have the same distribution for the treatment and control group.[19] We use the standardized differences test (Rosenbaum and Rubin, 1985) to examine the balance, defined as:

$$\text{STD}_{\text{diff}} = 100\frac{(\bar{x}_t - \bar{x}_c)}{\sqrt{\frac{s_t^2+s_c^2}{2}}}, \tag{2.5}$$

where $\hat{x}_t$ and $\hat{x}_c$ are the sample means for a particular covariate in the treated and control groups, respectively and s_t^2 and s_c^2 are the sample variances.

This standardized difference is computed for each covariate used in the matching procedure. One problem with this approach is that there is no formal criterion for how large this difference should be. Rosenbaum and Rubin (1985) propose that 20 be considered large. Table 2.5 shows the standardized differences both before and after our matching analysis (using Inverse Probability Weighting as explained below). The results are calculated from the formula above and dividing the outcome by 100. Table 2.5 shows that our preferred matching estimator has standardized differences of the important variables all lower than 0.2,[20] considered small in the literature. Table 2.5 shows the standardized differences for grant employment as treatment. The other standardized (Tables **??**-**??**) differences exhibit similar patterns.

Figure 2.3 shows the distribution of both the treated and non-treated students. We

[17]The AP Physics B, discontinued in 2014, is the equivalent of an introductory Physics college course. It was, later on, replaces by AP Physics 1 and 2. The AP Physics C: Mechanics test studies Newtonian mechanics, while the AP Physics C: Electricity and Magnetism test studies electricity and magnetism.

[18]The AP Calculus AB is the equivalent of the first Calculus course taken in college and includes topics such as limits, derivatives, definite integrals, and the Fundamental Theorem of Calculus. In comparison, the AP Calculus BC (CALBC) is a more advanced test and in addition to the topics included in the AP Calculus AB test it includes topics such as sequences and series.

[19]The balancing property is not equivalent to the CIA and vice-versa (Smith and Todd, 2005).

[20]This threshold was calculated by dividing 20 by 100.

can visually see the significant overlap between the two populations. Given this, we can proceed to estimate the ATET for our two outcomes and three treatments considered.

2.4 Results

Table 2.6 shows the estimates for the impact of grant employment on both general and STEM graduation rates. The estimates from Table 2.6 show an impact of about 10.1 percentage points of grant employment on STEM graduation rates and an impact of 5.5 percentage points on graduation rates. The nearest neighbor with replacement results with one neighbor are shown in Table 2.7 and they are very similar to the IPW results, which is reassuring. For the remainder of the paper, we only focus on the IPW results, since this similarity is maintained for the other tables as well.

We also divide grant employment based on how its research intensity, as explained in the Data Section. The results from Table 2.6 show the matching estimation results for the sample containing research-intensive jobs. The estimates show that research jobs and all grant employment have a similar effect on graduation rates, which could be caused by the fact that they both reduce the financial burden of a college education. Unfortunately, we do not have hourly wages for these jobs to be able to explicitly compare the reduction in the cost of attending college associated with each type of job. Another interesting result is that research jobs have a positive impact of 13.6 percentage points on STEM graduation rates, much higher than the effect of all grant employment on STEM graduation rates. These results suggest that while all types of employment on a federally funded grant improve both graduation and STEM graduation rates, the type of job held also matters and future research should also take this into account.

We are also interested in the heterogeneity of the treatment effect. When breaking down the analysis based on the gender of the students, we can see from Table 2.6 that women benefit more from research-intensive jobs than men. While the gender of the student doesn't seem to create a big divide for graduation rates based on the type of job the student held, it makes does matter for STEM graduation rates. When converting these changes into percent changes, we obtain a 25 percent higher effect for women than for men from having a research job on STEM graduation rates. This result can be tied to the stereotype threat literature, which states that women might not be able to act in accordance to their abilities in fear they might reinforce the negative stereotypes associated with their identity (Steele and Aronson, 1995). In this context, having a research-intensive job might help women succeed in STEM more by giving them more confidence in their ability to be a STEM major. Another potential cause of these results could be that the selection

problem is worse for women, issue which we address in the Sensitivity Analysis subsection. Another caveat is that, even though women benefit more from research experience than men, giving an extra grant to a female student rather than a male student reduces the number of STEM graduates, which is at variance with the goal of increasing STEM majors.

Another result of interest is the heterogeneous effect of research experience based on the race of the student (also shown in Table 2.6). Grant employment increases STEM graduation the most for Hispanic students and the least for African-American students. When examining only research-intensive positions, we can see that research jobs help white students the most with STEM graduation and Hispanic students the least. Table 2.6 also shows that the group with the lowest graduation rates, black students, benefits the most from grant employment. Hispanic students, also a group with low graduation rates, also benefit highly from being employed on a federally funded grant.

Table 2.6 also shows that grant employment increases graduation rates for students with Pell grants by 7.2 percentage points and it increases STEM graduation rates by 6.3 percentage points. To compare these effects with the effects for the full sample, we measure the percent change in graduation rates when compared to the control group mean. Thus, we conclude that having a job increases graduation rates for Pell grant recipients by more than for the full sample, but the opposite is true for STEM graduation. The fact that we get a different result for STEM graduation suggests that, although having a job decreases the cost of education, there might be other factors involved in getting a STEM graduation that we are not accounting for.

2.4.1 Inverse-probability-weighted regression adjustment estimator

We employ one last matching estimator for our analysis to check the robustness of our results to the estimator choice. The inverse-probability-weighted regression adjustment (IPWRA) estimator uses inverse probability weighting (IPW) weights when performing regression adjustment. It combines models for both the outcome and the treatment status. Wooldridge (2010) shows that this estimator is doubly robust meaning that the estimates of the effects will be consistent if either the treatment model or the outcome model, but not both, are misspecified.[21] Table 2.7 shows the results using the IPWRA estimator for grant employment. The results are very similar to the results obtained with the IPW estimator and they show an average treatment of the treated effect of grant employment on STEM graduation of 10.4 percentage points and on general graduation of 5.5 percentage points. For research intensive jobs, the results show an effect of 5.6 percentage points and of 14.0

[21] We could not find any studies that test the double robust property when both of the models are misspecified.

percentage points on STEM graduation rates.

2.4.2 Sensitivity analysis

A natural question arises regarding on how robust our estimates are to different sources of biases. One source of bias that could arise is due to the timing of research employment. Since the students who are doing research later in their academic career are also more likely to graduate, we are concerned that our estimates might be biased. As noted before, the majority of the students considered (70 percent) have their first grant employment opportunity in their first year of college. We also remove all instances of grant employment in the students' senior year since these students are very likely to graduate.

Another issue is that arises is that, by using a propensity matching technique, the selection problem is not necessarily fixed. The credibility of the matching procedure used relies entirely on the conditional independence assumption, which assumes that we observe all the variables that have an impact on research experience and graduation rates. We are interested in the sensitivity of our estimates with respect to confounding factors, i.e., unobserved variables that might influence both assignment to research experience and the likelihood of graduating in STEM.

While not possible to estimate the magnitude of selection bias without experimental data, it is possible to check the sensitivity of the estimated results with respect to deviations from the conditional independence assumption (Aakvik, 2001). Thus, we are interested in how unobserved covariates that affect both our treatments and our outcomes would alter our conclusions. Such unobserved variables could be motivation, future career aspirations, or any other factor that affects both the probability of taking part in undergraduate research experience, as well as the provability of graduation in general or in STEM). For example, a student who shows more motivation towards his or her studies might have a higher change of getting hired to work with a faculty member, and also might have a higher change of graduating from college. This section informs us how our results would change if we had such unobservable characteristics that would affect both our treatments and outcomes.

As a first step, we assume that the relationship between treatment and observable is related to the relationship between treatment and unobservables (Altonji, Elder and Taber, 2005). Our study includes a broad set of variables that cover socioeconomic characteristics, as well as data on the courses the students take and their high explanatory power suggests that the observable characteristics could provide useful information about the unobservable characteristics.

In our analysis, we control for pre-interest in STEM fields by including AP tests and

intent to major in STEM. Given that AP tests are highly predictive of research experience, we re-run our propensity score matching procedure by excluding AP tests, as shown in Table 2.8. The base estimates are the IPW estimates including AP tests, while the sensitivity analysis estimates are the ones excluding AP tests. The results show that excluding AP tests increases our estimates of the ATET effect of all types of grant employment on graduation rates from 5.5 percentage points to 5.8 percentage points. We get a similar increase in the ATET effect of all grant employment on STEM graduation rates, where by excluding the AP tests controls, our IPW estimates increase from 10.1 percentage points to 11 percentage points.

In general, these results suggest that if there were other factors such as motivation, that were equally as important in determining both employment and the outcomes in question as having an AP test, the exclusion of these other factors would bias upward our estimated effects, but by a relatively small magnitude.

Another approach to take when examining the extent to which our estimates are sensitive to biases is to calculate bounds to inform us of the degree to which the unobserved variables affect selection into treatment Rosenbaum (2002). With the notation used before, we have our binary outcome variable Y, our binary treatment T and a covariate vector X. Following the model in Aakvik (2001) and Becker and Caliendo (2007), we define probability of receiving treatment $\pi_i = Pr(x_i, u_i) = Pr(T_i = 1|x_i, u_i) = F(\beta x_i + \gamma u_i)$, where x_i are the observed characteristics for person i, while u_i are the unobserved characteristics. Here, γ is the effect that unobserved characteristics have on the treatment. In the case of no hidden bias, γ is zero, but in the presence of hidden bias, individuals with the same exact observed characteristics will have different probabilities of receiving treatment.

Assuming a matched pair of people i and j, and F a logistic regression, the odds that the individuals receive treatment are $\frac{\pi_i}{1-\pi_i}$ and $\frac{\pi_j}{1-\pi_j}$, respectively and the odds ratio is:

$$\frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}} = \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)} = \frac{exp(\beta x_i + \gamma u_i)}{exp(\beta x_j + \gamma u_j)} \tag{2.6}$$

The matching procedure implies that $x_i = x_j$, so we have:

$$\frac{exp(\beta x_i + \gamma u_i)}{exp(\beta x_j + \gamma u_j)} = exp(\gamma(u_i - u_j)) \tag{2.7}$$

When the odds ratio is one, there is no hidden bias. This happens when either the unobserved variables are the same ($u_i = u_j$), or when the unobserved variables do not influence selection into treatment ($\gamma = 0$). Assuming that the unobserved variable is a binary variable, we get from Rosenbaum (2002) the following bounds for the odds ratio:

$$\frac{1}{e^{\gamma}} \leq \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)} \leq e^{\gamma} \tag{2.8}$$

If $e^{\gamma} = 1$, then both individuals i and j have the same probability of being assigned the treatment. Here, e^{γ} measures the amount of departure from the case where we have no hidden bias (Rosenbaum, 2002). The equation above states that if two individuals differ in terms of unobservables, then their probability of receiving treatment depends on γ and the difference in u.

Thus, for a fixed $e^{\gamma} \geq 1$ and $u \in \{0, 1\}$, it can be shown that the test statistic Q_{MH} can be bounded by two known distributions (Rosenbaum, 2002).[22] In the case when $e^{\gamma} = 1$, the bounds are equal to the value of the test statistic. For values of e^{γ} greater than 1, the upper and lower bounds diverge as a consequence of unobserved selection bias creating uncertainty about the test statistics.

There are two STATA commands that have been developed for calculating these bounds: one for continuous outcomes-rbounds (DiPrete and Gangl, 2004), and one for binary outcomes-mhbounds (Becker and Caliendo, 2007). Since this paper deals with dichotomous outcomes, we employ the latter STATA command which uses the Mantel and Haenszel test statistic. The test can be used to test for no treatment effect both within different strata of the sample as well as a weighted average between the strata. In our example, we use a weighted average of the strata.

Following Aakvik (2001), we introduce additional notation in order to be able to apply this test. The outcome y is observed for both treatment and control groups, which under the null-hypothesis has a hypergeometric distribution. Furthermore, we define N_{1s} to be the number of treated individuals in stratum S and N_{0s} to be the number of untreated individuals in stratum s, so that $N_S = N_{1s} + N_{0s}$. In addition, Y_{1s} is the number of successful participants and Y_{0s} is the number of unsuccessful participants, which gives Y_s as the number of total successes in stratum s. With this notation in mind, the test statistic Q_{MH} is:[23]

[22] This test statistics can only be used after a matching procedure is performed since the individuals in the treatment and control groups have to be very similar to each other.

[23] This test statistic follows asymptotically the standard normal distribution.

$$
\begin{aligned}
Q_{MH} &= \frac{\mid Y_1 - \sum_{s=1}^{S} E(Y_{1s}) \mid -0.5}{\sqrt{\sum_{s=1}^{S} Var(Y_{1s})}} \\
&= \frac{\mid Y_1 - \sum_{s=1}^{S} \left(\frac{N_{1s} Y_S}{N_S} \right) \mid -0.5}{\sqrt{\sum_{s=1}^{S} \frac{N_{1s} N_{0s} Y_S (N_S - Y_S)}{N_S^2 (N_s - 1)}}}
\end{aligned}
\tag{2.9}
$$

We denote Q^{+}_{MH} to be the test statistic in the case where we overestimated the treatment effect and Q^{-}_{MH} to be the test statistic in the case where we underestimated the treatment effect. Thus, Q^{+}_{MH} statistic adjusts the MH test statistic downward for positive unobserved selection, while the Q^{-}_{MH} statistic adjusts the MH statistic upwards for negative unobserved bias. Then we have:

$$
Q^{+}_{MH} = \frac{|Y_1 - \sum_{s=1}^{S} \tilde{E}^{+}_{S}| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(\tilde{E}^{+}_{S})}} \tag{2.10}
$$

and

$$
Q^{-}_{MH} = \frac{|Y_1 - \sum_{s=1}^{S} \tilde{E}^{-}_{S}| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(\tilde{E}^{-}_{S})}} \tag{2.11}
$$

where $\tilde{E}_S$ and $Var(\tilde{E}_S)$ are the large-sample approximations to the expectation and variance of the number of participants in treatment.

To implement this procedure, we re-estimate the treatment effects using one-nearest-neighbor matching,[24] for different values of γ.The sensitivity analysis performed informs us how biases might influence our estimates, but does not inform us if biases exist.

[24]The estimates for the treatment effect using this procedure are slightly different from the estimates in the results section, due to the use of a different matching algorithm.

Table 2.2 shows that the students are positively selected into treatment based on observed characteristics. More specifically, students with higher high school GPAs, students who took AP science courses and students who have a higher ACT composite score are more likely to be treated. Positive observed selection into treatment does not imply positive unobserved selection into treatment, but it does inform us that some unobserved factors would confound our estimated effect.Furthermore, given that our estimated effect of research experience is positive, we are worried about overestimating our treatment effect. Therefore, we are only interested in the bias related to overestimation of the treatment effect.

Tables 2.9-2.12 show the sensitivity of the test statistic for e^{γ}, as well as for the test statistic $\Gamma = e^{\gamma} = 1$, the case with no hidden bias. All p-values are based on one-sided significance tests. The first column in the table represents $\Gamma = e^{\gamma} \geq 1$ for which the sensitivity analysis is carried out.

For the analysis of all grant employment on graduation rates, Table 2.9 shows that the test statistic becomes not significant at the 5 percent level when the relative odds of working on any federally funded grant with a faculty member are 3. Thus, we would need an unobserved variable u that increases treatment by 30 percent to make the relationship non-significant at a 5 percent significance level.

In order to interpret these results, we can compare the estimates from our sensitivity analysis with the results from logit models predicting treatment, the ones we estimated for our propensity score analysis. To put our sensitivity analysis results into perspective, Table 2.3 shows that the movement from not being eligible for a Pell grant to being eligible shifts the relative odds ratio by 1.4. To completely get rid of the effect of all grant employment on graduation rates, an unobserved variable would have to have almost as large of an effect as Pell grant eligibility, net of all the control variables we already include. Thus, this test provides evidence that a high amount of selection into unobservables is needed to eliminate the treatment effects.

The treatment effect stays positive for values of $\Gamma \leq 1.4$, after which point it becomes negative due to a large positive unobserved characteristics. Furthermore, we see from the same table that for values of $\Gamma \geq 1.6$, the test statistic becomes significant again at the standard 5 percent level.

Table 2.10 shows the sensitivity analysis results for the effect of research-intensive jobs on graduation rates. The effect at $\Gamma = 1$ is significant and stays significant until $\Gamma = 1.2$, when it is not even significant at 10 percent significance level. For $\Gamma = 1.2$, two students with the same observable characteristics differ in the odds of participating in research experience by a factor of 1.2, which is a very large number considering we have already ad-

justed for important student characteristics. Thus, the unobservable characteristics would have to increase the probability of receiving a research intensive job by 20 percent in order for the effects of research-intensive grant employment on graduation rates to stop being significant. This corresponds to a significant negative treatment effect, caused by large positive unobserved characteristics. Table 2.9 also shows that $\Gamma = 1.35$ is the point where the treatment effect changes signs, going from positive to negative In addition to this, the test statistic becomes significant again at 5 percent level once Γ exceeds 1.45, but the treatment effect becomes negative.

We again perform a back-of-the-envelope calculation to see which covariates from our propensity score regressions produce coefficients similar to the ones from the sensitivity analysis. In a similar way, we see from Table 2.4 that the "weakest" unobservable characteristics we would have to have in order to render the effect of research intensive jobs on graduation rates as not significant would have to be at least as large as the effect of taking the AP Biology test. Again, the unobservable characteristics would have to be variables that are not already included in our analysis.

Tables 2.11-2.12 present the Mantel-Haenszel bounds for the case where we consider STEM graduation as the outcome. For both tables, even the relative odds of getting treatment are 1.5, the test statistic p^{+}_{MH} is still statistically significant at a 5 percent significance level. When considering all grant employment as the treatment variable, Table 2.11 shows that the test statistic p^{+}_{MH} is statistically significant at 5 significance level for all values considered, except for the values for Γ between 1.55 and 1.8. Thus, for values of the odds ratio between 1.55 and 1.8, our inferences would become not significant at 5 percent significance level. In addition to this, for $\Gamma = 1.7$, the treatment effect changes signs, going from positive to negative.

Table 2.12 shows that our estimates of the effect of research-intensive jobs on STEM graduation would become statistically not significant at a 5 percent significance level for values of Γ between 1.7 and 2. The treatment effect at $\Gamma = 1.85$, although insignificant, remains positive. For values of $\Gamma \geq 1.85$, the treatment effect becomes negative.

These results are reassuring that while we cannot rule out the possibility of hidden biases, we can reassure ourselves that the unobservable characteristics would have to be quite large to make our results not significant. In both cases, it is quite unlikely to have unobservable characteristics that would switch the relative odds of receiving treatment by 1.55, and 1.7, respectively, given that switching the race or the gender of the student explains a much smaller change in the odds ratio. The results for considering STEM graduation rate as the outcome are less sensitive to unobservable characteristics than the results for general graduation rate as the outcome.

Mantel-Haenszel bounds for all the other subsamples are provided in Tables **??**-**??**. One way to read our results is by looking at the smallest value of Γ for which the effect of our treatment variables stops being statistically significant at 5 percent significance level, which we define Γ'. We also define Γ'' as the highest value of $\Gamma = e^{\gamma}$ for which the test statistic p^{+}_{MH} is still not significant at a 5 percent significance level. Thus, in the Table 2.12 from before, Γ' is equal to 1.7 and Γ'' is equal to 2.

The results are fairly robust to the possible presence of unobservable characteristics. For example, Table **??** shows that for the effect of all grant employment on graduation rates for female students to go away, we would need an unobserved variable that would multiply the odds of treatment by 1.25. As explained in the analysis above, this effect is fairly large given that it is net of all the controls we included. In general, the tables suggest that the estimates for Black and Hispanic students have larger intervals [Γ',Γ''] where the test statistic is not significant at 5 percent level. However, all of the subgroups considered are partially robust to selection bias. This implies that our estimates for these two groups are more sensitive to nonobservable characteristics, a result due in part to the lower representation of these groups in the overall student population. Another result is that the subgroup analysis is less robust to unobservable for general graduation rates than it is for STEM graduation rates. Thus, we need to be more careful when interpreting the estimates of our matching procedure when considering general graduation rates.

2.5 Conclusion and future research

The factors that impact persistence in STEM are of importance to the society due to the role of STEM graduates in technological advancement. This paper provides insights into the policy implications of research productivity by analyzing the role of grant employment on persistence in STEM. We use a unique dataset that combines administrative student transcript data with longitudinal administrative data on research funding at a public research institution. This innovative data allows us to track all employment and courses history for the students who attended this large university over 2001-2014.

Using this data, we quantify the impact of undergraduate employment, defined as having been employed on a federally funded grant at a large public university, on persistence in STEM. Using our preferred IPW estimator, we find a positive and significant impact of approximately 10.1 percentage points of grant employment on STEM graduation rates and an impact of approximately 5.5 percentage points on general graduation rates.

In addition, this study evaluates the effect of grant employment on different subgroups. We find that undergraduate employment helps narrow the gender, financial gaps, and

racial gaps. We also find that grant employment helps narrow the female-male STEM graduation gap. Furthermore, having worked with a faculty member helps Hispanic and Black students persist in college.

Finally, this study provides estimates of research employment for all types of grant employment, as well as more research-intensive positions. In particular, we divide research employment by the amount of research intensity. We define research jobs as the jobs that are related to the student's science career, based on the job description of the university's HR department. Our findings show that research jobs increase STEM graduation rates, and this effect is higher for female students than male students. For students with more financial constraints, such as the students receiving Pell grants, we find that working with a faculty member helps increase their graduation rates significantly.

Given the overall positive effects of research experience, we suggest the implementation of policies that would increase the research opportunities of undergraduate students at academic institutions. Furthermore, more employment opportunities should be available to female students and minority students, who are shown to benefit greatly from this type of employment. The costs of making more undergraduate research opportunities available should also be considered. In the future, we plan to extend our analysis to take into account the gender and race composition of the research teams that the students are part of. We envision extending our current research to investigate research collaborations in more depth, given that we have access to rich information on all the federally funded grants and their recipients. We also plan to investigate the effects of undergraduate research on longer-term student outcomes, such as graduate school attendance and labor market outcomes.

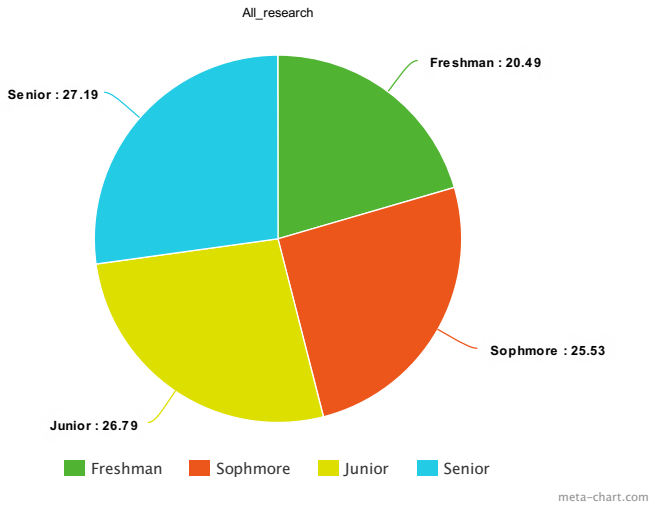
2.6 Tables and figures

Table 2.1: Literature review on undergraduate research experience

Study	Data	Type of research experience
Hurtado et al. (2009)	Focus groups (four research universities)	Structured science research programs
Barlow and Villarejo (2004)	University of California, Davis	Biology Undergraduate Scholars Program
Bauer and Bennett (2003)	Survey from University of Delaware	Undergraduate Research Program
Chang et al. (2014)	The Freshman Survey and College Senior Survey	Grants from NIH and NSF
Seymour et al. (2004)	Student interviews from four liberal arts colleges	Undergraduate research experience
Hunter, Laursen and Seymour (2007)	Ethnographic study at four liberal arts colleges	Summer science undergraduate research experiences
Kardash (2000)	Midwestern, Carnegie Research I university	Undergraduate science research
Russell, Hancock and McCullough (2007)	Web-based surveys	Undergraduate research opportunity (NSF)
Kim, Rhoades and Woodard Jr (2003)	22 public research universities	R&D expenditures from NSF's CASPAR data
Gregerman et al. (1998)	University of Michigan	Undergraduate Research Opportunity Program
Pender et al. (2010)	University of Maryland Baltimore County	Meyerhoff Scholarship Program
Hathaway, Nagda and Gregerman (2002)	Survey from University of Michigan	Undergraduate Research Opportunity Program
Lopatto (2004)	Online survey of undergraduates from 41 institutions	Summer undergraduate research programs
Zydney et al. (2002)	Survey at the University of Delaware	Undergraduate science research experience

Figure 2.1: Timing of research experience

(a) All research experience



(b) First time research experience

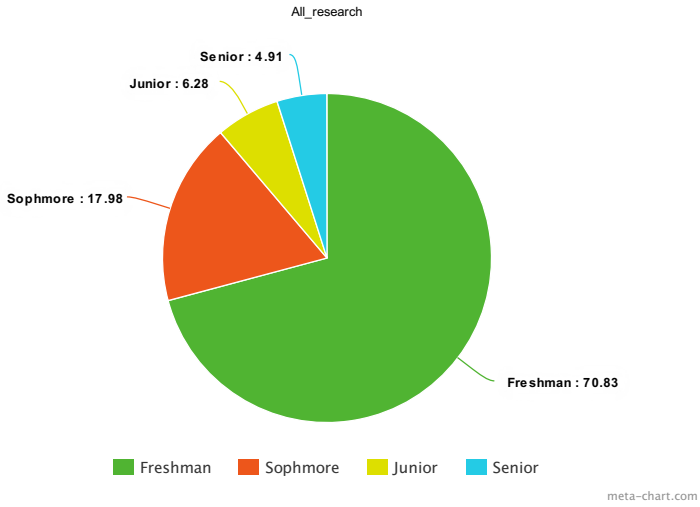


Figure 2.2: Number of months employed

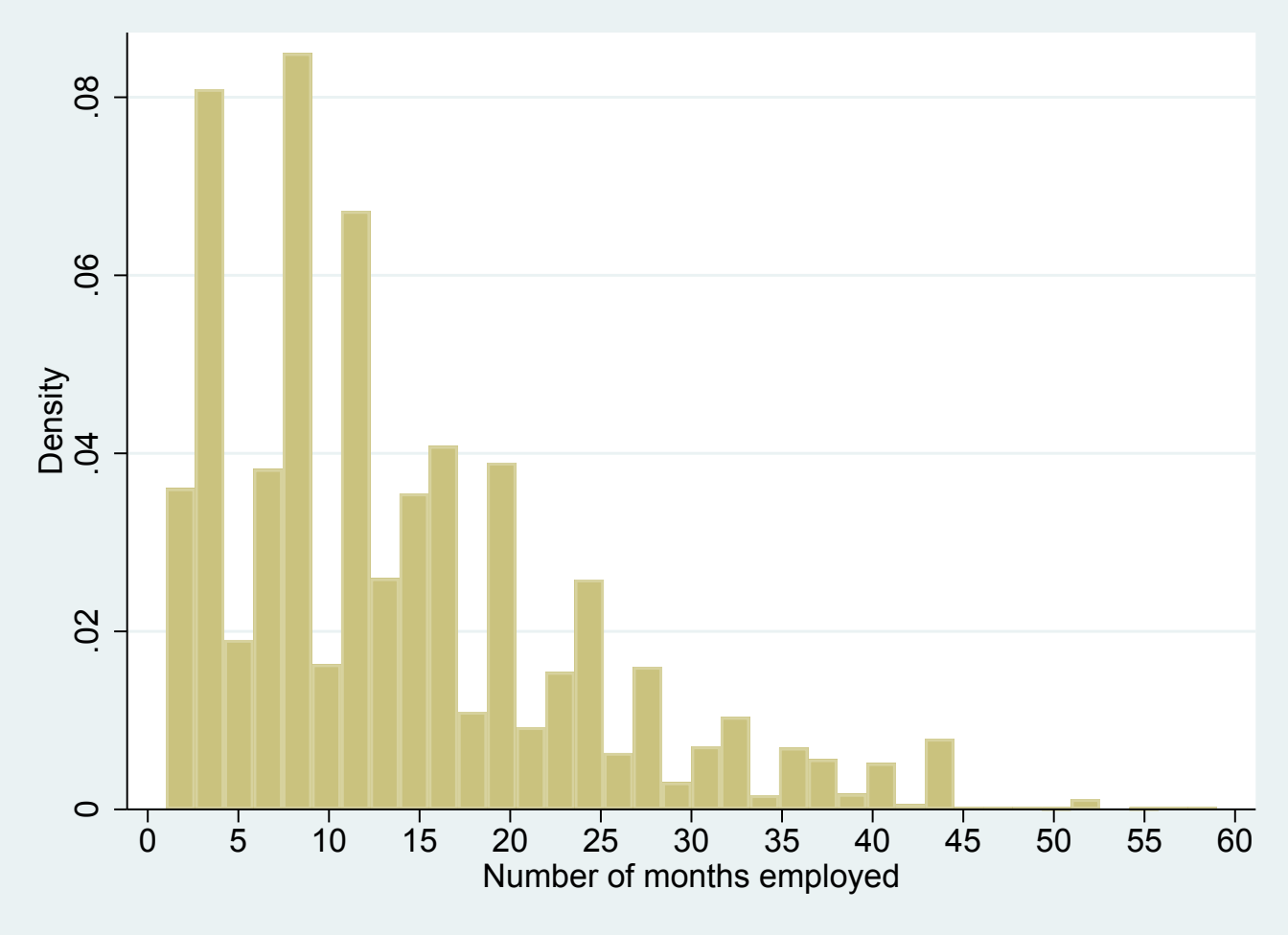


Table 2.2: Summary Statistics

	Full Sample		All jobs		Research jobs	
	Mean	SD	Mean	SD	Mean	SD
Female	0.529	0.499	0.594	0.491	0.568	0.495
Black	0.064	0.246	0.086	0.280	0.061	0.240
Hispanic	0.053	0.226	0.050	0.218	0.046	0.210
Asian	0.117	0.322	0.163	0.370	0.173	0.379
Other race	0.035	0.184	0.039	0.194	0.040	0.198
In state	0.645	0.479	0.771	0.420	0.780	0.415
High school GPA	3.723	0.292	3.781	0.251	3.801	0.232
International student	0.036	0.187	0.024	0.154	0.026	0.161
Pell grant	0.195	0.396	0.300	0.437	0.257	0.437
ACT composite score	28.10	3.445	28.34	3.351	28,71	3.35
Graduate	0.806	0.395	0.852	0.355	0.863	0.344
Graduate STEM	0.205	0.404	0.355	0.479	0.418	0.493
Graduate social science	0.285	0.451	0.223	0.416	0.177	0.382
AP BY score	0.626	1.467	0.919	1.741	1.057	1.844
AP CALAB score	1.140	1.816	1.351	1.940	1.477	1.997
AP CALBC score	0.453	1.307	0.609	1.503	0.694	1.593
AP CH score	0.490	1.260	0.758	1.534	0.863	1.620
AP CSA score	0.045	0.423	0.057	0.480	0.062	0.499
AP CSAB score	0.017	0.248	0.018	0.244	0.019	0.259
AP PHYSB score	0.166	0.747	0.209	0.864	0.244	0.938
AP PHYSE score	0.074	0.513	0.099	0.595	0.112	0.636
AP PHYSM score	0.201	0.853	0.303	1.042	0.348	1.115
AP STAT score	0.329	1.087	0.321	1.095	0.353	1.150
AP CALAB	0.253	0.435	0.302	0.459	0.329	0.470
AP CALBC	0.102	0.303	0.137	0.343	0.155	0.362
AP CALSB	0.109	0.312	0.145	0.352	0.165	0.371
AP BY	0.143	0.350	0.208	0.406	0.237	0.425
AP CH	0.112	0.316	0.175	0.380	0.200	0.400
AP CSA	0.010	0.100	0.013	0.114	0.014	0.118
AP CSAB	0.003	0.0611	0.004	0.064	0.004	0.066
AP PHYSB	0.039	0.195	0.048	0.215	0.057	0.233
AP PHYSE	0.015	0.122	0.020	0.142	0.022	0.150
AP PHYSM	0.045	0.209	0.068	0.253	0.079	0.270
AP STAT	0.078	0.268	0.074	0.263	0.081	0.274
Observations	35720	.	3653	.	2896	.

Notes: Five-year graduation rates reported.
Both AP scores and an indicator for taking AP tests are included.
Science AP tests included: Biology (BY), Chemistry (CH), Physics (Physics B (PHYSB), Physics C: Electricity and Magnetism (PHYSE), Physics C: Mechanics (PHYSM)), Computer Science (Computer Science A (CSA), Computer Science AB (CSAB)), Statistics (STAT), and Calculus (Calculus AB (CALAB), Calculus BC (CALBC)).

Table 2.3: Estimates of propensity score with any grant employment as the outcome, logit model

VARIABLES	Selection	Marginal effects
Female	0.195***	0.032***
	(0.023)	(0.003)
Black	0.154*	0.026*
	(0.063)	(0.010)
Hispanic	-0.028	-0.004
	(0.067)	(0.011)
Asian	0.181***	0.030***
	(0.041)	(0.007)
Other race	-0.023	-0.003
	(0.078)	(0.013)
In state	0.291***	0.048***
	(0.022)	(0.003)
HS GPA	0.047***	0.008***
	(0.008)	(0.001)
International student	0.073	0.012
	(0.059)	(0.010)
Pell grant	0.338***	0.056***
	(0.022)	(0.003)
ACT composite score	-0.042**	-0.007**
	(0.014)	(0.002)
ACT composite sq.	0.001***	0.0001***
	(0.000)	(0.000)
AP CALAB	0.063**	0.010**
	(0.033)	(0.003)
AP CALBC	0.048	0.008
	(0.033)	(0.005)
AP BY	0.144***	0.024***
	(0.025)	(0.004)
AP CH	0.123***	0.020***
	(0.028)	(0.004)
AP CSA	0.046	0.007
	(0.040)	(0.005)
AP CSAB	-0.003	-0.000
	(0.148)	(0.024)
AP PHYSB	0.097*	0.0163*
	(0.046)	(0.007)
AP PHYSE	-0.048	-0.008
	(0.081)	(0.013)
AP PHYSM	0.145**	0.024**
	(0.049)	(0.008)
AP STAT	-0.058	-0.009
	(0.035)	(0.006)
Constant	-1.829***	
	(0.195)	
Observations	35,720	35,720
Pseudo R-squared	0.058	0.058

Notes: All regressions contain race-gender interaction terms.
Selection equation estimated using a logistic regression. Marginal effects estimated at the mean of observable characteristics.
Standard errors in parentheses.
*** Statistical significance at the 1 percent level.
** Statistical significance at the 5 percent level.
* Statistical significance at the 10 percent level.

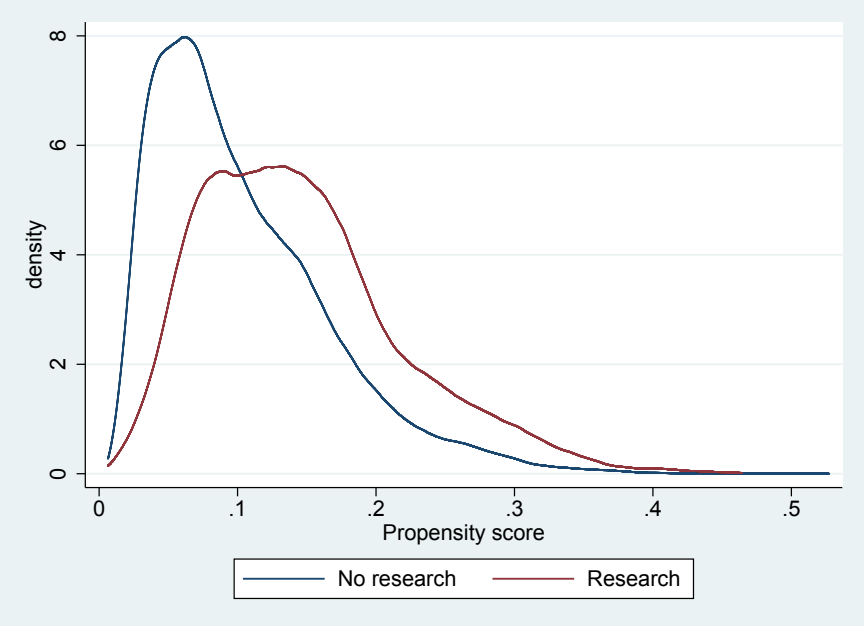
Table 2.4: Estimates of propensity score with research employment as the outcome, logit model

VARIABLES	Selection	Marginal effects
Female	0.160***	0.022***
	(0.024)	(0.003)
Other race female	0.126	0.017
	(0.106)	(0.014)
Black	0.032	0.004
	(0.074)	(0.010)
Hispanic	-0.0161	-0.002
	(0.0730)	(0.010)
Asian	0.176***	0.024***
	(0.043)	(0.006)
Other race	-0.050	-0.007
	(0.080)	(0.011)
In state	0.304***	0.042***
	(0.024)	(0.003)
HS GPA	0.057***	0.008***
	(0.008)	(0.001)
International student	0.144*	0.020*
	(0.063)	(0.008)
Pell grant	0.221***	0.031***
	(0.024)	(0.003)
ACT composite	-0.032	-0.004
	(0.017)	(0.025)
ACT composite score sq.	0.001**	-0.000**
	(0.000)	(0.000)
AP CALAB	0.074*	0.010*
	(0.024)	(0.004)
AP CALBC	0.050	0.007
	(0.034)	(0.004)
AP BY	0.182***	0.025***
	(0.026)	(0.003)
AP CH	0.140***	0.096***
	(0.029)	(0.004)
AP CSA	0.007	0.001
	(0.096)	(0.012)
AP CSAB	-0.045	-0.006
	(0.156)	(0.021)
AP PHYSB	0.143**	0.020**
	(0.047)	(0.006)
AP PHYSE	-0.068	-0.009
	(0.083)	(0.011)
AP PHYSM	0.164**	0.022**
	(0.050)	(0.007)
AP STAT	-0.035	-0.005
	(0.037)	(0.005)
Constant	-2.224***	
	(0.235)	
Observations	35,720	35,720
Pseudo R-squared	0.070	0.070

Notes: All regressions contain race-gender interaction terms.
Selection equation estimated using a logistic regression. Marginal effects estimated at the mean of observable characteristics.
Standard errors in parentheses.
*** Statistical significance at the 1 percent level.
** Statistical significance at the 5 percent level.
* Statistical significance at the 10 percent level.

Figure 2.3: Kernel density of probability of getting the treatment

(a) All grant employment



(b) Research jobs

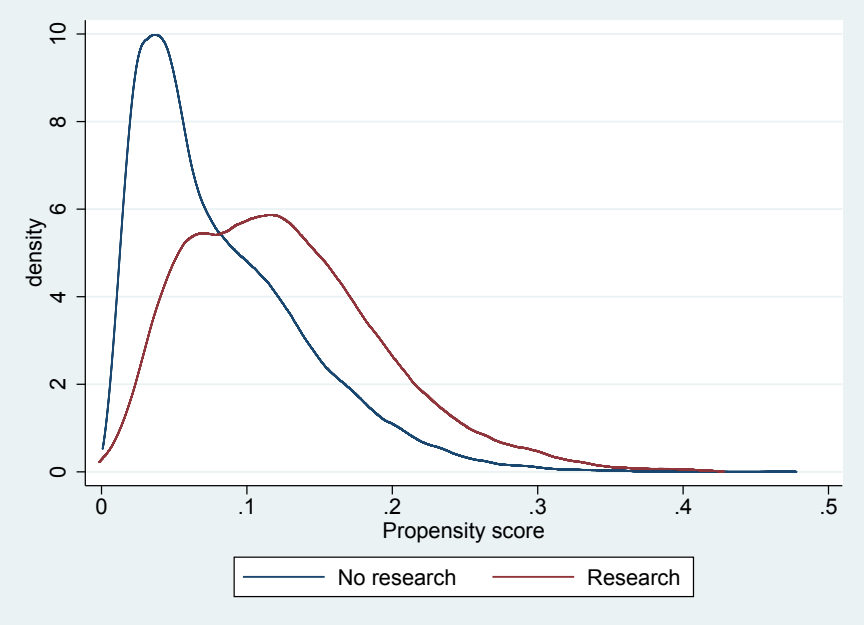


Table 2.5: Standardized differences for grant employment as treatment and graduation as outcome, inverse probability weighting

	Std diff Raw	Std diff Weighted	Var ratio Raw	Var ratio Weighted
Female	.147	-.004	.966	1.001
Black female	.117	-.005	1.648	.978
Hispanic female	.037	.001	1.230	1.008
Asian female	.137	-.005	1.570	.984
Other race female	.055	-.006	1.446	.962
Black	.090	-.008	1.348	.975
Hispanic	-.018	.003	.928	1.012
Asian	.148	-.005	1.371	.990
Other race	.024	-.003	1.125	.984
In state	.309	-.010	.758	1.014
Intent STEM	.395	-.008	.864	1.007
HS GPA	.208	-.007	.963	1.015
International student	-.077	.003	.654	1.020
Pell grant	.275	-.013	1.403	.988
Pell grant * female	.270	-.009	1.760	.985
ACT composite score	.078	.001	1.088	.978
ACT composite score sq.	.090	.002	1.108	1.018
AP CALAB	.121	-.000	1.131	.999
AP CALBC	.119	-.001	1.331	.996
AP BY	.191	.001	1.402	1.002
AP CH	.201	-.009	1.532	.983
AP CSA	.031	.000	1.341	1.008
AP CSAB	.006	-.000	1.106	.998
AP PHYSB	.050	.001	1.255	1.006
AP PHYSE	.046	-.005	1.410	.962
AP PHYSM	.111	-.002	1.552	.992
AP STAT	-.014	.002	.956	1.007

Notes: Standardized differences between the treatment and comparison groups, calculated based on the formula from Equation 2.5 divided by 100.

Table 2.6: IPW estimation results

		Graduation (percentage points)	STEM graduation (percentage points)
Overall	All jobs	0.055***	0.101***
		(0.006)	(0.008)
	Research jobs	0.056***	0.136***
		(0.006)	(0.009)
Female	All jobs	0.048***	0.0653***
		(0.008)	(0.010)
	Research jobs	0.049***	0.102***
		(0.009)	(0.012)
Male	All jobs	0.065***	0.147***
		(0.010)	(0.013)
	Research jobs	0.067***	0.176***
		(0.010)	(0.013)
White	All jobs	0.045***	0.108***
		(0.007)	(0.010)
	Research jobs	0.049***	0.146***
		(0.008)	(0.011)
Black	All jobs	0.101***	0.018
		(0.027)	(0.026)
	Research jobs	0.083**	0.075**
		(0.035)	(0.037)
Hispanic	All jobs	0.107***	0.094***
		(0.031)	(0.033)
	Research jobs	0.093***	0.084**
		(0.035)	(0.040)
Asian	All jobs	0.046***	0.109***
		(0.016)	(0.021)
	Research jobs	0.054***	0.126***
		(0.017)	(0.024)
Pell grant	All jobs	0.072***	0.063***
		(0.013)	(0.015)
	Research jobs	0.080***	0.110***
		(0.015)	(0.019)

Notes: Five-year graduation rates reported.
ATET results shown using the IPW estimator. All jobs refers to all types of federal grant employment, while research jobs refers to more research oriented types of federal grant employment.

Table 2.7: Nearest neighbor (1) and IPWRA estimation results

		Graduation (percentage points)	STEM graduation (percentage points)
IPW	All jobs	0.055***	0.101***
		(0.006)	(0.008)
	Research jobs	0.056***	0.136***
		(0.006)	(0.009)
NN(1)	All jobs	0.052***	0.101***
		(0.008)	(0.010)
	Research jobs	0.049***	0.134***
		(0.009)	(0.012)
IPWRA	All jobs	0.055***	0.104***
		(0.006)	(0.00808)
	Research jobs	0.056***	0.140***
		(0.006)	(0.009)

Notes: Five-year graduation rates reported.
The column base estimates presents results with all the controls included. They are the same results using the IPW estimator from Table 3.4. The column sensitivity analysis presents results using the same controls, excluding AP test controls, using the IPW estimator.

Table 2.8: Sensitivity analysis of the IPW estimation results

		Graduation (percentage points)		STEM graduation (percentage points)	
		Base estimates	Sensitivity analysis	Base estimates	Sensitivity analysis
Overall	All jobs	0.055***	0.058***	0.101***	0.110***
		(0.006)	(0.006)	(0.008)	(0.008)
	Research jobs	0.056***	0.060***	0.136***	0.148***
		(0.006)	(0.006)	(0.009)	(0.009)
Female	All jobs	0.048***	0.051***	0.0653***	0.073***
		(0.008)	(0.008)	(0.010)	(0.010)
	Research jobs	0.049***	0.051***	0.102***	0.114***
		(0.009)	(0.009)	(0.012)	(0.012)
Male	All jobs	0.065***	0.069***	0.147***	0.158***
		(0.010)	(0.010)	(0.013)	(0.013)
	Research jobs	0.067***	0.072***	0.176***	0.189***
		(0.010)	(0.010)	(0.013)	(0.014)
White	All jobs	0.045***	0.047***	0.108***	0.117***
		(0.007)	(0.007)	(0.010)	(0.010)
	Research jobs	0.049***	0.052***	0.146***	0.159***
		(0.008)	(0.008)	(0.011)	(0.011)
Black	All jobs	0.101***	0.099***	0.018	0.028
		(0.027)	(0.027)	(0.026)	(0.027)
	Research jobs	0.083**	0.086**	0.075**	0.093**
		(0.035)	(0.035)	(0.037)	(0.038)
Hispanic	All jobs	0.107***	0.107***	0.094***	0.108***
		(0.031)	(0.031)	(0.033)	(0.034)
	Research jobs	0.093***	0.094***	0.084**	0.097**
		(0.035)	(0.035)	(0.040)	(0.041)
Asian	All jobs	0.046***	0.049***	0.109***	0.121***
		(0.016)	(0.016)	(0.021)	(0.022)
	Research jobs	0.054***	0.056***	0.126***	0.136***
		(0.017)	(0.017)	(0.024)	(0.0244)
Pell grant	All jobs	0.072***	0.074***	0.063***	0.067***
		(0.013)	(0.013)	(0.015)	(0.015)
	Research jobs	0.080***	0.084***	0.110***	0.119***
		(0.015)	(0.015)	(0.019)	(0.019)

Notes: Five-year graduation rates reported.
the column base estimates presents results with all the controls included. They are the same results using the IPW estimator from Table 3.4. The column sensitivity analysis presents results using the same controls, excluding AP test controls, using the IPW estimator.

Table 2.9: Sensitivity analysis using Mantel-Haenszel bounds for grant employment and graduation

$\Gamma = e^{\gamma}$	Q^{+}_{MH}	Q^{-}_{MH}	p^{+}_{MH}	p^{-}_{MH}
1.00	5.630	5.630	<0.001	<0.001
1.05	4.844	6.420	<0.001	<0.001
1.10	4.094	7.17	<0.001	<0.001
1.15	3.379	7.898	<0.001	<0.001
1.20	2.695	8.593	.003	<0.001
1.25	2.040	9.261	.020	<0.001
1.30	1.411	9.905	.079	<0.001
1.35	.806	10.527	.209	<0.001
1.40	.224	11.129	.411	<0.001
1.45	.275	11.711	.391	<0.001
1.50	.818	12.275	.206	<0.001
1.55	1.343	12.823	.089	<0.001
1.60	1.852	13.356	.031	<0.001

$\Gamma = e^{\gamma}$: odds of differential assignment due to unobserved factors
Q^{+}_{MH}: Mantel-Haenszel statistic (assumption: overestimation of treatment effect)
Q^{-}_{MH}: Mantel-Haenszel statistic (assumption: underestimation of treatment effect)
p^{+}_{MH}: significance level (assumption: overestimation of treatment effect)
p^{-}_{MH}: significance level (assumption: underestimation of treatment effect)

Table 2.10: Sensitivity analysis using Mantel-Haenszel bounds for research job and graduation

$\Gamma = e^{\gamma}$	Q^{+}_{MH}	Q^{-}_{MH}	p^{+}_{MH}	p^{-}_{MH}
1.00	3.773	3.773	<0.001	<0.001
1.05	3.102	4.446	<0.001	<0.001
1.10	2.462	5.090	.006	<0.001
1.15	1.852	5.706	.031	<0.001
1.20	1.268	6.297	.102	<0.001
1.25	.709	6.866	.239	<0.001
1.30	.172	7.415	.431	<0.001
1.35	.271	7.944	.392	<0.001
1.40	.770	8.456	.220	<0.001
1.45	1.250	8.951	.105	<0.001
1.50	1.715	9.431	.043	<0.001

$\Gamma = e^{\gamma}$: odds of differential assignment due to unobserved factors
Q^{+}_{MH}: Mantel-Haenszel statistic (assumption: overestimation of treatment effect)
Q^{-}_{MH}: Mantel-Haenszel statistic (assumption: underestimation of treatment effect)
p^{+}_{MH}: significance level (assumption: overestimation of treatment effect)
p^{-}_{MH}: significance level (assumption: underestimation of treatment effect)

Table 2.11: Sensitivity analysis using Mantel-Haenszel bounds for grant employment and STEM graduation

$\Gamma = e^{\gamma}$	Q^{+}_{MH}	Q^{-}_{MH}	p^{+}_{MH}	p^{-}_{MH}
1.00	9.430	9.430	<0.001	<0.001
1.05	8.508	10.355	<0.001	<0.001
1.10	7.629	11.238	<0.001	<0.001
1.15	6.791	12.083	<0.001	<0.001
1.20	5.990	12.895	<0.001	<0.001
1.25	5.222	13.675	<0.001	<0.001
1.30	4.485	14.426	<0.001	<0.001
1.35	3.776	15.150	<0.001	<0.001
1.40	3.094	15.850	<0.001	<0.001
1.45	2.436	16.526	.007	<0.001
1.50	1.800	17.181	.035	<0.001
1.55	1.185	17.816	.117	<0.001
1.60	.590	18.433	.277	<0.001
1.65	.0136	19.032	.494	<0.001
1.70	.492	19.614	.311	<0.001
1.75	1.035	20.181	.150	<0.001
1.80	1.563	20.733	.589	<0.001
1.85	2.077	21.271	.018	<0.001

$\Gamma = e^{\gamma}$: odds of differential assignment due to unobserved factors
Q^{+}_{MH}: Mantel-Haenszel statistic (assumption: overestimation of treatment effect)
Q^{-}_{MH}: Mantel-Haenszel statistic (assumption: underestimation of treatment effect)
p^{+}_{MH}: significance level (assumption: overestimation of treatment effect)
p^{-}_{MH}: significance level (assumption: underestimation of treatment effect)

Table 2.12: Sensitivity analysis using Mantel-Haenszel bounds for research job and STEM graduation

$\Gamma = e^{\gamma}$	Q^{+}_{MH}	Q^{-}_{MH}	p^{+}_{MH}	p^{-}_{MH}
1.00	10.402	10.402	<0.001	<0.001
1.05	9.555	11.253	<0.001	<0.001
1.10	8.748	12.065	<0.001	<0.001
1.15	7.978	12.842	<0.001	<0.001
1.20	7.242	13.588	<0.001	<0.001
1.25	6.536	14.305	<0.001	<0.001
1.30	5.859	14.996	<0.001	<0.001
1.35	5.208	15.662	<0.001	<0.001
1.40	4.582	16.305	<0.001	<0.001
1.45	3.977	16.927	<0.001	<0.001
1.50	3.394	17.529	<0.001	<0.001
1.55	2.830	18.113	.002	<0.001
1.60	2.284	18.680	.0111	<0.001
1.65	1.755	19.230	.039	<0.001
1.70	1.242	19.765	.107	<0.001
1.75	.744	20.286	.228	<0.001
1.80	.260	20.793	.397	<0.001
1.85	.152	21.288	.439	<0.001
1.90	.610	21.770	.270	<0.001
1.95	1.057	22.241	.145	<0.001
2.00	1.492	22.701	.067	<0.001
2.05	1.916	23.150	.027	<0.001

$\Gamma = e^{\gamma}$: odds of differential assignment due to unobserved factors
Q^{+}_{MH}: Mantel-Haenszel statistic (assumption: overestimation of treatment effect)
Q^{-}_{MH}: Mantel-Haenszel statistic (assumption: underestimation of treatment effect)
p^{+}_{MH}: significance level (assumption: overestimation of treatment effect)
p^{-}_{MH}: significance level (assumption: underestimation of treatment effect)

CHAPTER III

Gender and persistence in STEM

From a work with Margaret Levenstein and Jason Owen-Smith

Abstract

Although women have surpassed men in college persistence, female students remain much less likely to major in STEM fields. This paper uses administrative student data from a large public university to study the effects of students' socio-demographic and academic characteristics on the necessary and weakly sequential stages to achieve a STEM degree: taking a STEM course in the first year, declaring a STEM major, and graduating with a STEM major. Using a model similar to that of Heckman and Smith (2004), we find that female students are 7.7 percentage points less likely than male students to graduate with a STEM degree. These results are driven by the male students declaring a STEM major at a higher rate than female students. Once a STEM major was declared, no statistically significant differences exist in the probability of graduating in STEM between the two genders. Our findings suggest that exploring the different mechanisms affecting the differential propensities of male and female students to major in STEM could inform policies aimed at reducing the under-representation of women in STEM fields.

JEL-Classification: I20, I23, J16, J15
Keywords: Higher education, teaching assistants, STEM persistence

3.1 Introduction

While the number of total STEM (Science, Technology, Engineering, and Math) college degrees has increased considerably in the past two decades, there are still strikingly large gender gaps. Even though women have outnumbered men in terms of college enrollment and attainment of bachelor's and higher level degrees (Snyder and Dillow, 2015; Goldin, 2006; Turner and Bowen, 1999), STEM fields still exhibit a gender gap in degree attainment (DiPrete and Buchmann, 2013). Furthermore, even in STEM fields such as life sciences, where women have outnumbered men in terms of bachelor degrees received (Mann and DiPrete, 2013), there still exist large differences in grades, class participation, and views on being knowledgeable about the subject (Eddy, Brownell and Wenderoth, 2014; Grunspan et al., 2016).

This under-representation of women in STEM careers further contributes to the gender pay gap since STEM fields pay on average higher salaries (Beede et al., 2011). This gender pay gap was estimated to be between $.18$ and $.21$ in terms of average hourly earnings for full-time American workers in the mid-2000s (Blau and Kahn, 2016). One solution offered to ameliorate the gender pay gap problem is to attract more female students into STEM fields.

Previous studies have tried to track the trajectories of female and male students in STEM over time in an attempt to get a better understanding of the factors that drive women away from STEM. The gender gap in math and science grades was found to be non-existent at the primary and secondary school level (Xie and Shauman, 2003; Kenney-Benson et al., 2006). Furthermore, even though previously existent, gender gaps in high school have now practically disappeared as female students are now as equally likely as male students to take calculus and science classes (NSB, 2014). This gender equality at high school level changes drastically as the students enter postsecondary education, where male students are more likely than female students to choose a science or mathematics major (Hill, Corbett and St Rose, 2010).

The low percentage of female students who select STEM majors (Sax, 2008) has motivated researchers to study the STEM pipeline at the undergraduate level. One can think of the STEM pipeline as the trajectory of students in STEM from primary school all the way to a STEM career. The absence of women in STEM in tertiary education has been metaphorically referred to as a "leaky pipeline" (Blickenstaff, 2005). One reason why we might encounter these "leaks" could be due to some students having an initial interest in STEM not ending up declaring a STEM major. Another event causing a "leak" is the case where a student declares a STEM major but does not end up graduating in STEM, he/she

either graduates in a non-STEM major or drops out. These "leaks" are also more common among female and underrepresented minority students. Xie and Shauman (2003) argue that using the "pipeline" approach is too simplistic since it assumes a linear trajectory of achieving a STEM career. While we recognize that the "pipeline" approach has some shortfalls, we believe that this approach is appropriate for our study given our narrow focus on STEM persistence, combined with considering all the required steps to graduate with a STEM major at the university studied. Furthermore, we consider a broader definition of declaring a STEM major, by taking into account all the possible majors a student declared in their undergraduate career, and not just the first one declared.

In this paper, we analyze the stages of persistence in a STEM major to examine which stages have the most impact on persistence and further determine which subgroups are the most affected at each stage. We use a model similar to the one used by Heckman and Smith (2004) for participation in a job market training program and define persistence in STEM as a process that requires the following sequential stages: taking a STEM class in the first year of college, declaring a STEM major, and graduating with a STEM degree. While we focus on the differential persistence rates based on gender, we also examine other demographic characteristics of the students that could affect persistence, such as race and financial aid status. Our findings show that female students are 7.7 percentage points less likely to graduate in STEM than the male students, but this result disappears when we only compare female and male students who have declared a STEM major. We also find that minority students have a higher propensity to take STEM courses in the first year, but that they are not more likely than white students to graduate in STEM. The students most likely to graduate in STEM are also the ones with higher ACT composite scores and high school GPAs.

In an attempt to get more insight into these findings, we also simulate the effect of changing one student characteristic on the overall probability of completing a STEM degree and on each stage in the process. Unsurprisingly, we find once again that female students are less likely than male students to complete a STEM degree, an effect that mainly stems from the fact that female students are much less likely than male students to declare a STEM major conditional on having taken a STEM class in their Freshman year. Regarding race, we find that underrepresented minority students are less likely to graduate in STEM, result consistent with previous studies (Herrera and Hurtado, 2011; Goyette and Xie, 1999). One interesting result is that even though blacks and Hispanics both diverge from their initial interest in STEM, they do so at different stages of persistence- graduation and STEM declaration, respectively.

Thus, previous suggestions to improve the representativeness of URMs in STEM that

focus on promoting academic confidence are still valid, but they need to be tailored according to the stages of persistence where the students drop out. Thus, solutions suggested by the previous literature, such as undergraduate research programs (Grandy, 1998; Graham et al., 2013; Ward, Bennett and Bauer, 2003), supplemental instruction (Villarejo and Barlow, 2007), tutoring (Perna et al., 2009), and career support and development (MacLachlan, 2006) are still relevant, but need to be further developed.

The rest of the paper is organized as follows. In Section 3.2 we describe the previous literature. In Section 3.5 we describe our dataset, and in Section 3.4 we present our theoretical model. Section 3.3 talks about the institutional background, while Section 3.6 shows our empirical model. Section 3.7 presents the main results. The final section summarizes our findings and discusses policy implications.

3.2 Literature review

Various explanations have been put forth to explain the gender differences in STEM for both persistence, as well as career choice.[1] One of the explanations for the lack of women in STEM is the fact that female students express lower confidence related to their mathematical abilities, even in the case where they perform better than the male students (Sax, 1994; Vogt, Hocevar and Hagedorn, 2007; Correll, 2001; , n.d.; Micari, Pazos and Hartmann, 2007). Cech et al. (2011) use panel data from four academic institutions from Massachusetts and reiterate the previous findings that the undergraduate STEM persistence gap is mostly due to the women's lack of confidence.

Some papers have attempted to relate the lower confidence of STEM female students to differences in innate abilities of men and women (Ceci and Williams, 2009, 2010; Murphy, Steele and Gross, 2007). However, recent studies have shown that these differences are mostly due to factors that could potentially be altered, such as cultures influences (Guiso et al., 2008; Nollenberger, Rodríguez-Planas and Sevilla, 2016; Pope and Sydnor, 2010), a fixed mindset that favors men (Good, Rattan and Dweck, 2012), an unfriendly STEM climate (Meinholdt and Murray, 1999; Hill, Corbett and St Rose, 2010), and existent gender-specific stereotypes (Nguyen and Ryan, 2008; Aronson and McGlone, 2009; Seymour and Hewitt, 1997; Johnson et al., 2012).

Another cause of the dearth of women in STEM majors is the lack of women in STEM careers (Urry, 2015). One solution suggested has been increasing the number of female faculty members in STEM, which in turn could act as role models for the female students. This body of research suggest that female faculty members can act as role models in STEM

[1] See Blickenstaff (2005) and Xie and Shauman (2003) for summaries of these explanations.

fields and impact both persistence (Griffith, 2014; Bettinger and Long, 2005; Price, 2010; Robst, Keil and Russo, 1998) and choice of major (Canes and Rosen, 1995; Rothstein, 1995; Carrell, Page and West, 2010; Qian, Zafar and Xie, 2009; Rask and Bailey, 2002). The majority of previous studies has found a positive correlation between the gender of the instructor and the students' STEM persistence, with female students being more likely to pursue a STEM major (Rask and Bailey, 2002; Qian, Zafar and Xie, 2009; Canes and Rosen, 1995; Rothstein, 1995; Carrell, Page and West, 2010), get higher grades (Hoffmann and Oreopoulos, 2009; Griffith, 2014) and persist in STEM majors (Hoffmann and Oreopoulos, 2009; Carrell, Page and West, 2010). Although promising, these findings are at odds with other studies that find no effect or negative effects on either persistence or STEM grades (Ehrenberg, Goldhaber and Brewer, 1995; Griffith, 2010; Price, 2010).

While the mechanisms might still be unclear, there is a common consensus that women are less likely than men to persist in STEM majors (Griffith, 2010), and this holds even for high performing female students (Bettinger, 2010). This phenomenon creates a vicious cycle where women in engineering are discriminated against more than the men (Vogt, Hocevar and Hagedorn, 2007), and in turn, women in STEM fields feel discouraged to attend large introductory courses (Johnson, 2007) perhaps being apprehensive of ending up in a STEM career.

Our paper aims to contribute to the literature in various ways. First of all, we present a conceptual framework previously used in the job training literature to examine the stages of persistence in STEM. While our main focus is on differences in persistence based on gender, we provide additional analysis based on the race and financial status of the students. Second of all, we use a rich administrative student data from a public Midwestern institution that allows us to control for pre-interest in STEM, measured by AP tests, and interest to choose a STEM major in college. This rich dataset also contains information about high school grade point average, as well as the whole history of courses taken by the students in their undergraduate career.

Finally, this study provides estimates for the effect of changing one characteristic on the overall probability of completing a STEM degree and on each stage in the process. This exercise allows us to identify the stages in persistence that are the most relevant for each subgroup considered and informs our suggestions for future policies to improve diversity in STEM.

3.3 Institutional background

The public Midwestern institution that we study is composed of various colleges, with the main one (approximately 60% enrollment) being the College of Arts and Sciences. The College of Engineering is the second largest college. The academic calendar is comprised of four terms, with Fall and Winter being the main ones and Spring and Summer the secondary ones.

Students in the College of Arts and Sciences typically declare a major during the second term of their sophomore year. In the College of Engineering, the earliest a student can declare a major is during their second term attending the university, and all students are urged to declare a major by the start of their third term enrolled in courses. Furthermore, students in the College of Engineering cannot register for courses in their fourth term enrolled unless they have declared a major.

To declare a major, the student is required to meet with the department advisor and make sure that he/she has fulfilled the requirements for declaring that particular major. The course requirements for each major vary, ranging from 24 to 48 credits at the 200-level and above. Each student must earn an overall GPA of minimum 2.0 in the courses taken for their major. Students can change their major at any time after the first major declaration, as long as they have the approval of the department advisor in the new major chosen. Furthermore, students can even complete a second major after they receive a degree in a first major. They can do so by registering as a non-degree candidate and taking the corresponding coursework for the completion of a second degree.[2] The students also cannot freely internally transfer between the different colleges at the university, or declare majors that do not belong in their specific college.

The university offers academic minors, typically requiring at least 15 credits of course work. Choosing a minor is optional and there is no upper limit to the number of minors a student can choose. In our analysis, we disregard the students' minors and only focus on their majors. Each student can elect multiple majors, but they must meet all the requirements for all the majors.

3.4 Conceptual framework

The framework we use for analyzing graduation with a STEM degree is based on the one developed by Heckman and Smith (2004). We decompose graduating with a STEM degree into three consecutive steps: taking a STEM course in the first year, declaring

[2] We remove all these instances from our analysis.

a STEM major and graduating with a STEM degree.[3] With these steps in mind, for each vector of controls x_i corresponding to student i, we define the probability of taking a STEM course in the first year as $Pr(STEM\_cls_i = 1|x_i)$, the probability of declaring a STEM major conditional on having taken a STEM class in the first year as $Pr(STEM\_maj_i = 1|STEM\_cls_i = 1, x_i)$, and the probability of graduating with a STEM major conditional on having declared a STEM major and having taken at least a STEM course in the first year as $Pr(grad\_STEM_i = 1|STEM\_maj_i = 1, STEM\_cls_i = 1, x_i)$. Thus, the probability of graduating in a STEM major conditional on x_i is:

$$\begin{aligned} Pr(grad\_STEM_i = 1|x_i) = Pr(&STEM\_cls_i = 1|x_i)* \\ &Pr(STEM\_maj_i = 1|STEM\_cls_i = 1, x_i)* \\ &Pr(grad\_STEM_i = 1|STEM\_cls_i = 1, STEM\_maj_i = 1, x_i) \end{aligned} \tag{3.1}$$

Furthermore, using the chain rule, we obtain the following:

$$\begin{aligned} \frac{\partial Pr(grad\_STEM_i = 1|x_i)}{\partial x_{ij}} = &\frac{\partial Pr(STEM\_cls_i = 1|x_i)}{\partial x_{ij}}* \\ &Pr(STEM\_maj_i = 1|STEM\_cls_i = 1, x_i)* \\ &Pr(grad\_STEM_i = 1|STEM\_cls_i = 1, STEM\_maj_i = 1, grad_i = 1, x_i) \\ &+ Pr(STEM\_cls_i = 1|x_i)* \\ &\frac{\partial Pr(STEM\_maj_i = 1|STEM\_cls_i = 1, x_i)}{\partial x_{ij}}* \\ &Pr(grad\_STEM_i = 1|STEM\_cls_i = 1, STEM\_maj_i = 1, grad_i = 1, x_i) \\ &+ Pr(STEM\_cls_i = 1) \\ &+ Pr(STEM\_cls_i = 1)* \\ &Pr(STEM\_maj_i = 1|STEM\_cls_i = 1, x_i)* \\ &\frac{\partial Pr(grad\_STEM_i = 1|STEM\_cls_i = 1, STEM\_maj_i = 1, grad_i = 1, x_i)}{\partial x_{ij}} \end{aligned} \tag{3.2}$$

We apply this framework in all of the next sections to analyze graduation with a STEM

[3]Since Bettinger (2010) finds that the proportion of first-year STEM courses is a good predictor of majoring in STEM, we consider the first step of STEM persistence the probability of taking a STEM course in the first year.

degree. Given this chain rule decomposition, we can now determine at which stage and in which direction certain characteristics of the students affect graduation in STEM. Based on this equation, we can decompose the effect of each characteristic x of the student on STEM graduation into smaller effects on the probability at each stage weighted by the remaining probabilities. In the case of binary variables, derivatives are replaced with finite changes.

3.5 Data

This section describes the data that we use for our analysis.

3.5.1 Data on student outcomes

We use administrative student data from a public Midwestern institution. The sample is restricted to undergraduate students who attended this large Midwestern public institution from Fall 2001 to Winter 2014 and were admitted as Freshmen. We remove any transfer students, whose course-taking behavior might vary due to past college experience. Focusing on Freshmen helps us identify the courses that students take in their first year of college and also the semester in which they declare a major. We calculate graduation rates based on a five-year window, starting from the first semester the student is enrolled for courses at this university. In order to allow students to graduate in 5 years, we restrict the sample to students who take courses before Fall 2010.

The administrative data offer a combination of socio-demographic information, pre-college experience, and course taking behavior. The data cover the basic demographic information and the entire course taking history of each student. We consider the following race categories: white, black, Hispanic, Asian, and other (native American, not indicated, Hawaiian and two or more). We define international students as students with their country of residency outside of the United States at the beginning of their first year of studies.

We have additional data on Advanced Placement (AP) exams and information about the last high school attended by the student. Since the analysis in this paper focuses on STEM outcomes, we focus on the science and math AP tests.[4] In addition to AP tests, we also control for high school grade point average (GPA), recalculated by the university on a 4.0 scale.[5] Additional controls included are the standardized test scores, with SAT

[4]The AP tests considered are: Biology, Chemistry, Physics (Physics B, Physics C: Electricity and Magnetism, and Physics C: Mechanics), Computer Science (Computer Science A and Computer Science AB), Statistics, and Calculus (Calculus AB and Calculus BC).

[5]A caveat is that before 2009, the university included only the courses taken in grades 9-11 for calculating the GPA. After 2009, the university considered all high school courses taken for all grades.

composite scores converted into ACT composite scores.[6]

Given that the data on parental education and income acquired from the admission office contain a very large number of missing observations (over 40 percent for parental income and over 20 percent for parental education) and the fact that we cannot use multiple imputation methods due to the non-randomness of the missing data, we use need-based grant eligibility as a proxy. We create a binary Pell grant variable that identifies students who have received a Pell grant, the largest of the need-based financial grants that assists low-income students.

The administrative data include detailed information on course-specific outcomes such as grades, course subject, registration status, number of credits earned, as well as student outcomes such as graduation, persistence, and degree obtained. We also collect information about intended major prior to attending college from three different sources: the Common Application, the SAT, and the ACT. The Common Application asks students to list their areas of interest in college, with no required upper bound for the answers provided. The SAT exam contains a questionnaire on the choice of major, allowing up to three answers. The ACT asks students to list the college major they plan to have, with only one answer allowed.

We denote as STEM all the fields thought to contribute to technological innovation (Xie, Fang and Shauman, 2015). Although there are various STEM definitions, we choose the definition designated by the U.S. Immigration and Customs Enforcement agency on April 2008 when the extension for the Optional Practical Training (OPT) was introduced.[7] We do not take into account the additions to the list of STEM degrees in 2011 and 2012 (when fields like psychology, agriculture, etc. were added to the STEM list).

3.5.2 Summary statistics

Table 3.1 presents descriptive statistics for our data. The main dataset has one observation per student. Our sample of undergraduate students who attended a large Midwestern public institution from Fall 2001 to Winter 2010 and admitted as Freshmen contains 35720 students. We further restrict the sample by removing the 48 students who students don't take a STEM class in the first year but declare a STEM major. This procedure leaves us with a total of 35627 students.

In our sample of 35672 remaining students, 53 percent are female students. The most

[6]We use the official ACT conversion table: http://www.act.org/aap/concordance/pdf/reference.pdf.

[7]The Optional Practical Training (OPT) is a period during which undergraduate and graduate students on a student visa are allowed to work for one year.

represented race is white, with 67 percent of the students belonging to this race. The second largest race is represented by Asians, making up 12 percent of the student population. Hispanics and blacks represent about 5, and respectively 6 percent of the student population. International students are about 4 percent of the full sample and 19 percent of all the students have received Pell related grants. Out of all the students, 64 percent are in state students.

Table 3.1 also presents the summary statistics broken down by gender. We can see that a higher proportion of international students and in-state students are male. Furthermore, male students also have a higher ACT composite score than the female students. Of interest are the variables identifying intent to major in STEM. We also see a difference of 15 percentage points in the intent to major in STEM between the two genders. Since intent to major in STEM is a binary variable, taking the value one if any of the majors listed is a STEM major, we are also interested in the fraction of majors listed as potential majors in college that were STEM majors. Breaking this fraction by genders, we can see that male students list a much higher fraction of STEM majors, as compared to female students. The summary statistics also show a higher propensity of male students to take a STEM class in the first year, declare a STEM major and graduate with a STEM degree. These statistics suggest that the differences in incoming characteristics could explain some of the gap in STEM persistence rates between the female and the male students.

3.5.3 At each stage conditional on the previous stages

Table 3.2 presents the summary statistics at each stage considered, conditional on the previous stages. There are two transitions that we are interested in: moving from taking at least one STEM class in the first year to declaring a STEM major and moving from declaring a STEM major to graduating with a STEM degree. We can see that 86 percent of students take at least one STEM class in the first year, with men being slightly more likely than women to take a STEM class in the first year (90 percent of all men versus 83 percent of all women take a STEM class in the first two semesters). Asian students are the most likely to take a STEM class in the first year (94 percent of them do). Most of the international students also take a STEM class in the first year, and about 87 percent of the students who receive Pell grants do so as well.

The table also shows that 36 percent of the students who take a STEM class in the first year declare a STEM major at some point in time and that men are more likely than women to declare a STEM major. Once they declare a STEM major, 77 percent of students graduate with a STEM degree and male and female students are equally likely to graduate in STEM. Furthermore, white students are most likely of all races to graduate in STEM

after they declare a STEM degree.

One important question to answer with this table is how important is the students' intent to major in STEM. The last column of Table 3.2 is informative of this matter and suggests that students who intend to major in STEM are more likely to take a STEM course in the first semester, more likely to declare a STEM major, and only slightly more likely to graduate in STEM.

3.6 Empirical model

In our empirical model, we focus on the characteristics that impact each stage of undergraduate STEM persistence: taking a STEM course in the first year, declaring a STEM major and graduating with a STEM degree. The main empirical model that we estimate consists of an ordinary least-square regression that takes on the following specification:

$$y_{it} = \beta_0 + \beta_1 \text{Female}_{\text{it}} + \beta_2 X_{it} + \rho_t + \epsilon_i \tag{3.3}$$

The outcome y_{it} is the probability of achieving a specific outcome at each one of the stages leading to graduation in STEM. Our first outcome is a binary variable for having taken any STEM course in the first year of attending college. The second outcome considered is the probability of ever declaring a STEM major, conditional on having taken a STEM class in the first year of college. The final outcome is the probability of graduating with a STEM degree, conditional on having declared a STEM major and having taken at least one STEM class in the first two semesters of undergraduate education. The coefficient β_1 captures the effect of being a female student on each stage of persistence in STEM.

Each regression contains indicators for demographic characteristics (race, gender-race interactions), in-state status (as measured by the address of the student at the time of enrollment), international student status and cohort fixed effects (ρ_t). We also include interest in STEM as a control since previous studies show that an interest in STEM at high school level is correlated with STEM degree completion (Maltese and Tai, 2011, 2010). Because high school academic performance has also been shown to be highly correlated with the choice of major in college (Arcidiacono, Aucejo and Hotz, 2016; Kokkelenberg and Sinha, 2010; Rask, 2010; Ellington, 2006), we include high school GPA, standardized scores, and AP tests in our analysis. To account for the differences in STEM persistence based on socio-economic status (Schneider, Swanson and Riegle-Crumb, 1998; Miller and Kimmel, 2012; Hellerstein and Morrill, 2011), each regression contains indicators for being eligible for a Pell grant.

3.7 Results

This section presents the empirical results of our analysis. As a benchmark, we show the main results using the decomposition from the previous section that only contain gender controls in Table 3.3. These results suggest that female students are 17 percentage points less likely to graduate in STEM than their male counterparts, and that most of the effect comes from the fact that they are also 23 percentage points less likely than male students to declare a STEM major, after taking at least a STEM course in their first year.

Table 3.4 shows our results using the whole set of controls using ordinary least squares regressions. Each column of the Table 3.4 represent a different stage of persistence in STEM. The first and last columns both represent the last stage of persistence in STEM, which is STEM graduation rate, with the caveat that the first column is the unconditional outcome, while the last one is the outcome conditional on all the previous stages of persistence in STEM (taking a STEM course in the first year and declaring a STEM major).

A number of interesting findings emerge from our analysis. Without conditioning on the previous stages of STEM persistence, women are 10 percentage points less likely than men to graduate in STEM. Column 1 also suggest that black students are also less likely to graduate in STEM compared to white students. Students from the state where the university is located and students from outside the United States are also more likely to graduate with a STEM degree. In addition, STEM degree recipients also seem to be positively selected based on high school GPA, but not ACT composite scores.

To get a clearer image of the mechanisms that cause these results, we focus on columns 2-4. These additional results show that gender is a statistically significant predictor of the probability of being in the first two stages of STEM persistence: female students are 4 percentage points less likely to take a STEM class the first year of college and 14 percentage points less likely to declare a STEM major after taking at least a STEM class in the first year. Even after controlling for race, in-state status, high school grade point average, financial aid, ACT composite score and international student status, this effect stays negative and significant. Since there is barely any discrepancy between the rate at which female and male students take at least a STEM class in the first year, this result can be tied back to the initial 15 percentage points difference in the self-reported intent to major in STEM between the two genders. Surprisingly, once declaring a major, women in STEM are not significantly less likely than men to graduate in STEM.

Switching our attention to race, Table 3.4 shows that blacks, Hispanics, and Asians are all more likely than white students to take a STEM course in the first year. This result does not translate into higher probabilities of declaring a STEM major or graduating in STEM.

As a matter of fact, black students who declared a STEM major are less likely to graduate in STEM.

International students are more likely to take a STEM class the first year, declare a STEM major, and graduate in STEM. In-state students are more likely to take a STEM class in the first year, declare a STEM major, but less likely to graduate in STEM. Somehow expected, students from more disadvantaged backgrounds are less likely to graduate in STEM once they declare a STEM major. Surprisingly though, they are more likely to declare a STEM major after they take STEM courses in the first year.

Intent to major in STEM proves to be a strong predictor of entry into a STEM major, with students with pre-college interest in STEM being more likely to take a STEM course in their first year of college and also declare a STEM major. In addition, high school GPA is also positively correlated with being on the STEM pipeline. While ACT composite score also is a strong indicator of taking a STEM course in the first semester, as ACT score increases the effect lessens. Thus, a high ACT composite score predicts a high rate of taking a STEM course in the first year, but it does not predict the rate of STEM declaration or that of graduation in STEM.

We also simulate the effect of changing one characteristic on the overall probability of completing a STEM degree and on each intermediate step in the process. In particular, we examine the effect of changing a student's gender, race, ACT composite score, high school GPA, and Pell grant recipient status. These decompositions provide us with unique information on how different student characteristics such as gender, race and financial status affect persistence in STEM majors. Some characteristics even have opposing effects at different stages of persistence and this analysis helps us shed light on the process of acquiring a degree in STEM based on different student characteristics. The results are shown in Tables 3.5 and 3.6.[8] The standard errors reported in the parenthesis are bootstrap standard errors from 5000 bootstrap iterations and they reflect variation based on the sample used for the simulation. In addition, the derivatives or finite differences reported are the averages of the individual derivatives or finite differences, and not the derivatives evaluated at the means of the characteristics.

Each column of the table corresponds to a term from equation 3.2. The second column corresponds to the left-hand side term of equation 3.2, which is the overall effect of a change in each characteristic x on the probability of obtaining a STEM degree. Columns three, five and seven contain the three different components that are part of the overall effect. These weighted x terms are the terms with the finite differences of x with respect

[8]We do not include as many controls in these regressions, due to the nature of our code, thus these results are not directly comparable to the previous ones.

to x_{ij} weighted by the probabilities of reaching prior and later stages. Thus, column three represents the first term of equation 3.2, column five the second and column seven represents the last term. The remaining columns, column four, six and eight, represent the percentage that each stage contributes to the overall effect. We calculate this term by dividing the weighted term to the overall effect and then multiplying the result by 100.

By decomposing the overall probability of STEM graduation using this method, we can identify the stage where each student falls out of the STEM pipeline. The decomposition results for gender and race are shown in Table 3.5, and the ones for ACT composite scores, high school GPA and Pell grant eligibility are shown in Table 3.6. The results in Table 3.5 show that women are 17.9 percentage points less likely than men to complete a STEM degree. Women are also 1.8 percentage points less likely to take a STEM class the first year and 15.5 percentage points less likely to declare a STEM major. Overall, female students have lower conditional probabilities at every stage. The dominant factor decreasing the overall probability of female students graduating with a STEM degree is a strong negative probability of declaring a STEM major.

When switching our attention to race, we can see that blacks and Hispanics are less likely to graduate in STEM, as compared to white students. Switching race from white to black, while keeping all the other characteristics constant, makes a student 0.3 percentage points less likely to graduate in STEM (result which is however, not statistically significant). In addition, switching the race of a student from white to Asian makes him/her 8.9 percentage points more likely to graduate in STEM. While the main effect for black students is coming from the lower rate of STEM graduation, the main effect for Hispanics comes from their lower propensity to take STEM courses in the first year. Blacks, Asians and students of other race are all more likely to declare a STEM major. These results are partially consistent with Arcidiacono, Aucejo and Spenner (2012) who find that minorities are more likely to declare a STEM major, as compared to their white counterparts.

Table 3.6 presents the results for the other characteristics considered and it shows that switching Pell grant eligibility status from zero to one reduces the probability of graduating with a STEM degree by 0.3 percentage points. Furthermore, the probability of graduating in STEM increases monotonically in ACT composite score and high school GPA, consistent with previous studies that find that prior academic achievement (such as high school GPA and ACT/SAT scores) is one of the strongest predictors of persistence in STEM (Crisp, Nora and Taggart, 2009). Thus, moving from the lowest tercile to the middle and the high ones for both measures of ability increases the probability of graduating with a STEM degree. Given that we hold constant all the other characteristics of the students when doing this analysis, we can conclude that by targeting our policies at lower ability students, we could

be increasing substantially the number of students in STEM.

3.8 Conclusion and future research

Gender differences in STEM persistence are a very complex matter. In this study, we use a rich student administrative data from a large, public institution to provide evidence on the female students' trajectory on the STEM "pipeline". By considering the different consecutive stages of obtaining a STEM degree at undergraduate level, we are able to offer suggestions on policy interventions to improve the representation of women in STEM.

In our paper, we consider three important stages of persistence in STEM: taking at least one STEM course in the first two semesters of attending the university, declaring a STEM major at any point in time, and graduating with a STEM degree. Our estimates from a preliminary ordinary least-squared analysis suggest that women are less likely than men to be at all the stages of persistence in STEM. These results are consistent with the literature analyzing the gender gap in STEM persistence and finding that women "leak" out of the STEM pipeline (Gayles and Ampaw, 2014; Blickenstaff, 2005). In particular, we find that women are 4 percentage points less likely than men to take a STEM course in the first year, 14 percentage points less likely to declare a STEM but not significantly less likely than men to graduate in STEM (once they declare a STEM major). We also find that while minority students have a higher propensity to take STEM courses in the first year, they are not more likely than white students to graduate in STEM. As expected, the measures of ability are positively correlated with the probability of being on the STEM pipeline.

We also simulate the effect of changing one characteristic on the overall probability of completing a STEM degree and on each stage in the process. While we get similar results as in the previous analysis, we are also able to identify the effect of each stage of persistence on the final stage which would further enable us to propose relevant policies. Again, female students are less likely than male students to complete a STEM degree, effect that mainly stems from the fact that female students are much less likely than male students to declare a STEM degree. This result is consistent with studies have linked the underrepresentation of women in STEM to major choice in college (Turner and Bowen, 1999; Boudarbat and Montmarquette, 2009; Brown and Corcoran, 1997).

Our findings also point out that, while Hispanics and black students are less likely than white students to graduate in STEM, they fall off the STEM pipeline at different stages: Hispanics at the stage of declaring a major and blacks at the STEM graduation stage. These results informs our future policies to increase minorities in STEM and suggest the need to tailor the policies to different racial groups. Students from disadvantaged backgrounds fall

off the pipeline too, mostly due to the lower probability to graduate in STEM once they declare a major.

Even though this paper does not study the mechanisms behind these results, we include a discussion of the possible factors that could be driving these results, with a focus on gender. One possible explanation for women declaring STEM majors at a lower rate than men might be due to the grades women receive in STEM. Based on Arcidiacono (2004), students learn over time about their abilities in college and choose a major accordingly. Furthermore, Stinebrickner and Stinebrickner (2013) show that the main reason why students drop out of STEM is that they learn that their math and science abilities are lower than expected. Grades in introductory courses could be a strong signal for students learning about their abilities in college and they have been shown to have a lasting impact on the probability of persisting in the major (Ost, 2010). In addition, STEM field departments are generally known to be the most difficult grading ones in general. Koester, Grom and McKay (2016) find gendered differences in grades for large introductory STEM courses, with male students experiencing smaller grade penalties[9] than female students. This problem is also exacerbated by the fact that there is a different degree of sensitivity to grades for men and women. Female students have been shown to be more sensitive to grades in Economics courses (Dynan and Rouse, 1997; Rask and Tiefenthaler, 2008; Goldin, 2013) and physical sciences (Ost, 2010). In addition, Correll (2001) finds that grades have a more negative effect on persistence of female students than male students when transitioning from high school to college.

Another explanation for the lack of females in STEM is that the STEM environment could be driving the female students away. Zafar (2013) finds that female students chose majors depending on whether they enjoyed the coursework. Again, this issue can be connected to the chilly environment that STEM might provide, combined with the lack of female role models. If the policy objective is to improve female participation in STEM careers, one potential policy response is to increase the proportion of female STEM faculty, given the positive effects of female faculty on female students (Carrell, Page and West, 2010). In general, changing attitudes about stereotypes would offer the female students an opportunity to declare STEM majors at a higher rate. Previous research shows that female students' apprehension of math and science can be removed by lowering the stereotype threat (Spencer, Steele and Quinn, 1999). However, this intervention might not be enough to close the gender gap. In a study examining why women leave science and

[9]Grade penalty is defined as the difference between the grade in the class considered and GPA in all the other classes received by the student up to that point. This difference is referred to as a penalty to reflect the fact that in general students have lower grades in STEM courses as compared to other courses.

engineering fields, Hunt (2016) finds that women are more likely to leave the STEM fields because of salaries and promotion opportunities.

Future research is needed to understand the mechanisms behind women's lack of participation in STEM. In order to have a successful intervention to bridge the gender gap in STEM at the undergraduate level, we need to fully understand the reasons why female students are not as likely to declare STEM majors as their male counterparts. Examining the factors that help female students graduate in STEM is a necessary, but not sufficient step in keeping women in STEM careers. We should continue to focus on the differential impact of college experiences by gender in order to tease out the mechanisms that affect persistence, both in STEM and also in non-STEM majors.

3.9 Appendix tables and figures

Table 3.1: Summary statistics for male and female students

	Full Sample		Male		Female	
	Mean	SD	Mean	SD	Mean	SD
Female	0.53	0.50				
White	0.67	0.47	0.67	0.47	0.68	0.47
Black	0.06	0.25	0.05	0.23	0.074	0.26
Hispanic	0.05	0.23	0.05	0.23	0.051	0.22
Asian	0.12	0.32	0.12	0.32	0.12	0.32
Other race	0.03	0.18	0.03	0.18	0.035	0.18
Race missing	0.05	0.23	0.06	0.25	0.04	0.21
International student	0.03	0.19	0.04	0.21	0.02	0.16
Pell grant	0.19	0.40	0.19	0.39	0.20	0.40
ACT composite score	28.1	3.44	28.5	3.47	27.7	3.38
In state	0.64	0.48	0.62	0.49	0.67	0.47
Intent to major in STEM	0.52	0.50	0.59	0.49	0.45	0.50
Any STEM first year	0.86	0.35	0.90	0.30	0.83	0.38
Declared STEM major	0.28	0.45	0.40	0.49	0.17	0.38
Ever graduated with STEM degree	0.21	0.41	0.30	0.46	0.13	0.34
Observations	35672		16810		18862	

Table 3.2: Summary statistics at each stage conditional on the previous stage

		Full Sample	Male	Female	White	Black	Hispanic	Asian	Other race	In state	HS GPA	Int. Student	Pell grant	Intent STEM
STEM	Mean	0.86	0.9	0.83	0.85	0.86	0.87	0.94	0.86	0.88	0.77	0.93	0.87	0.96
first year	Obs	18372	16810	18862	24038	2305	1920	4182	1251	23006	6692	1297	6953	18372
Ever declared	Mean	0.32	0.44	0.21	0.32	0.23	0.26	0.41	0.32	0.34	0.23	0.5	0.32	0.5
STEM major	Obs	17562	15083	15634	20428	1979	1663	3920	1081	20296	5146	1200	6048	17562
Ever grad	Mean	0.74	0.74	0.74	0.76	0.57	0.68	0.75	0.72	0.74	0.67	0.77	0.67	0.75
in STEM	Obs	8829	6689	3294	6544	465	436	1619	349	6993	1208	594	1955	8829

Table 3.3: OLS results at each stage of persistence in STEM

VARIABLES	(1) Graduate STEM (unconditional)	(2) Any STEM first year (unconditional)	(3) Declare STEM (conditional)	(4) Graduate STEM (conditional)
Female	-0.17***	-0.07***	-0.23***	-0.00
	(0.00)	(0.00)	(0.01)	(0.01)
Constant	0.30***	0.90***	0.44***	0.74***
	(0.00)	(0.00)	(0.00)	(0.01)
Observations	35,672	35,672	30,717	9,983
R-squared	0.04	0.01	0.06	0.00
Adj. R-squared	0.0419	0.00972	0.0617	-9.16e-05

Notes: All specifications include year fixed effects and controls for AP science scores. Robust standard errors are reported in the parenthesis.
*** $p<0.001$, ** $p<0.01$, * $p<0.05$

Table 3.4: OLS results at each stage of persistence in STEM

VARIABLES	(1) Graduate STEM (unconditional)	(2) Any STEM first year (unconditional)	(3) Declare STEM (conditional)	(4) Graduate STEM (conditional)
Female	-0.10***	-0.04***	-0.14***	-0.00
	(0.00)	(0.00)	(0.00)	(0.01)
Black	-0.02*	0.04***	-0.02	-0.05*
	(0.01)	(0.01)	(0.01)	(0.02)
Hispanic	-0.00	0.04***	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.02)
Asian	-0.00	0.06***	0.00	-0.01
	(0.01)	(0.00)	(0.01)	(0.01)
Other race	0.00	0.02	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.02)
In state	0.03***	0.05***	0.05***	-0.02*
	(0.00)	(0.00)	(0.01)	(0.01)
HS GPA	0.03***	0.02***	0.02***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)
International student	0.21***	0.11***	0.25***	0.14***
	(0.01)	(0.01)	(0.01)	(0.02)
Pell grant	-0.00	0.00	0.03***	-0.05***
	(0.00)	(0.00)	(0.01)	(0.01)
ACT composite score	-0.01	0.02***	-0.01	0.00
	(0.00)	(0.00)	(0.00)	(0.01)
ACT composite sq.	0.00*	-0.00***	0.00*	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Intent to major in STEM	0.24***	0.16***	0.34***	0.02
	(0.00)	(0.00)	(0.00)	(0.01)
Constant	0.01	0.40***	0.06	0.42***
	(0.04)	(0.05)	(0.06)	(0.09)
Observations	35,672	35,672	30,717	9,983
R-squared	0.24	0.10	0.29	0.08

Notes: All specifications include year fixed effects and controls for AP science scores. Robust standard errors are reported in the parenthesis.
*** $p<0.001$, ** $p<0.01$, * $p<0.05$

Table 3.5: STEM degree completion simulation results (gender and race)

	Overall Effect	Weighted STEM first year term	% of Overall	Weighted STEM major term	% of Overall	Weighted STEM degree term	% of Overall
Change sex from male to:							
Female	-0.18	-0.015	8.553	-0.158	87.718	-0.002	1.143
	(0.004)	(0.001)	(0.473)	(0.004)	(2.063)	(0.002)	(1.237)
Change race from white to:							
Black	-0.003	0.009	-260.977	0.001	-22.146	-0.012	365.272
	(0.010)	(0.001)	(41.398)	(0.008)	(242.105)	(0.005)	(147.398)
Hispanic	-0.002	0.009	-480.92	-0.006	320.236	-0.007	389.898
	(0.010)	(0.001)	(77.921)	(0.008)	(420.809)	(0.005)	(266.074)
Asian	0.089	0.019	20.818	0.056	62.462	0.009	10.573
	(0.007)	(0.001)	(1.073)	(0.006)	(6.497)	(0.003)	(3.152)
Other race	0.020	0.006	27.930	0.018	90.895	-0.003	-17.375
	(0.012)	(0.002)	(10.170)	(0.010)	(49.636)	(0.005)	(27.389)

Notes: The estimates shown are calculated from Equation 3.2. Bootstrap standard errors using 5000 bootstrap repetitions appear in parenthesis.

Table 3.6: STEM degree completion simulation results (ACT, HS GPA and Pell grant)

	Overall Effect	Weighted STEM first year term	% of Overall	Weighted STEM major term	% of Overall	Weighted STEM degree term	% of Overall
Change ACT composite score from lowest tercile to:							
Second tercile	0.058	0.005	9.459	0.046	78.856	0.001	2.39
	(0.005)	(0.001)	(1.058)	(0.002)	(3.900)	(0.001)	(2.517)
Third tercile	0.104	0.01	9.742	0.096	91.896	0.003	2.629
	(0.006)	(0.001)	(1.016)	(0.005)	(4.768)	(0.003)	(2.742)
Change HS GPA from lowest quartile to:							
Second tercile	0.075	0.007	9.467	0.031	42.029	0.033	43.489
	(0.005)	(0.000)	(0.625)	(0.002)	(2.707)	(0.002)	(2.037)
Third tercile	0.148	0.013	8.696	0.065	43.968	0.052	35.374
	(0.007)	(0.001)	(0.523)	(0.004)	(2.912)	(0.002)	(1.384)
Change from no Pell grant to:							
Pell grant	-0.003	0.001	-24.816	0.016	-515.725	-0.018	583.04
	(0.006)	(0.001)	(31.589)	(0.005)	(149.670)	(0.003)	(91.770)

Notes: The estimates shown are calculated from Equation 3.2. Bootstrap standard errors using 5000 bootstrap repetitions appear in parenthesis.

APPENDIX A

Chapter I Supporting Material

A.1 Data Appendix

Table A.1: Evaluation items for overall quality category

Overall, the instructor was an excellent teacher.
Overall, the TA was an excellent teacher.
Overall, the lab instructor was an excellent teacher.

Table A.2: Evaluation items for TA effort category

The exams were returned in a reasonable amount of time.
Graded assignments (e.g. exams, papers) were returned in a reasonable amount of time.
The instructor was accessible to students outside of class.
The instructor handled questions well.
The TA handled questions well.
The instructor was open to contributions from all class members.
The lab instructor used techniques to foster class participation.
The instructor seemed well prepared for each class.
The instructor was well-prepared for each class.
The TA seemed well prepared for each class.
The instructor explained material clearly and understandably.
The instructor gave clear explanations.
The instructor presented material clearly in lectures/discussions.
The instructor delivered clear, organized explanations.
The TA gave clear and understandable explanations.
The lab instructor gave clear explanations.
The instructor used class time well.
The lab instructor used class time well.
The instructor helped me to understand the subject matter.
The instructor thoroughly understood the subject matter.
The instructor appeared to have a thorough knowledge of the subject.
The TA appeared to have a thorough knowledge of the subject.

Table A.3: Evaluation items for environment category

Students felt comfortable asking questions.
The instructor treated students with respect.
Grades were assigned fairly and impartially.
Grading was a fair assessment of my performance in this course.
The TA graded papers (exams, homework) fairly.
The instructor was concerned that we learn.
The instructor was willing to help students outside of class.
The instructor gave individual attention to students in the class.
The instructor was sensitive to student difficulty with course work.
The instructor motivated me to work hard.
The instructor set high standards for students.
The instructor made the course difficult enough to be stimulating.
The class meetings were stimulating and informative.
This course increased my desire to learn more about this subject in the future.
I can see myself furthering my education in this area.
I deepened my interest in the subject matter of this course.
I developed enthusiasm about the course material.
The instructor was accessible to students outside of class.
The instructor had regular office hours and was available at those hours.
The instructor was willing to help students outside of class.
The instructor suggested specific ways students could improve.
The instructor kept students informed of their progress.
The instructor told students when they had done especially well.
The instructor made the course interesting.
The instructor seemed to enjoy teaching.
The instructor was enthusiastic.
The instructor maintained an atmosphere of good feeling in class.
I was very satisfied with the educational experience this instructor provided.
I would take another course with this instructor.
The instructor was enthusiastic about the subject matter.
The instructor was friendly.
My teacher demonstrates a strong commitment to teaching.
My teacher is fair and impartial when dealing with me.
The instructor was confident and in control of the class.
Students' difficulty with the material was recognized.
The instructor showed a genuine concern for the students.
The instructor knows me by name.
The instructor suggested specific ways students could improve.
The instructor was skillful in observing student reactions.
The lab instructor kept students informed of their progress.
The lab instructor set high standards for students.
The lab instructor taught in a manner that served my needs as a student.
The instructor brought out the best in me as a student.
The instructor encouraged student participation in an equitable way.
The instructor made good use of examples and illustrations.
The instructor made me feel known as an individual in this course.
The instructor made the course interesting.
The instructor maintained an atmosphere of good feeling in class.
The instructor responded effectively to student difficulty in class.

Table A.4: Evaluation items for undergraduate student learning category

I learned a great deal from this course.
I gained a good understanding of concepts/principles in this field.
I deepened my interest in the subject matter of this course.
I developed the ability to communicate clearly about this subject.
I learned to apply principles from this course to new situations.
I learned a great deal in this laboratory.
I learned a great amount of substantive material.
I learned a great deal from this course.
I learned a great deal in this laboratory.
I learned to apply principles from this course to new situations.
I gained a good understanding of concepts/principles in this field.
I gained valuable experience working in teams in this course.
I increased my ability to analyze and interpret data.
I increased my ability to apply math and science knowledge to engineering problems.
I increased my ability to collect original data.
I increased my ability to design and conduct experiments.
I increased my ability to formulate, and solve engineering problems.
My confidence in my design abilities increased because of this course.
My oral communication skills improved because of this course.
My writing improved because of this course.
Course improved my ability to communicate technical information, designs, and analyses.

A.2 Introductory STEM courses

A.2.1 Mathematics

The university considered offers four Math sequences.[1] My analysis only consists of the courses that are part of the standard sequence since the other courses are taught by entirely by faculty members. The standard sequence is taken by undergraduate students who plan to major in sciences or engineering and it contains the following courses: Calculus I, Calculus II and Calculus III. Calculus I and II consist of only lectures, while Calculus III has both a lecture and a laboratory (see Table A.5). Out of the instructors teaching Calculus I, 64.3 percent are graduate students, while 68.5 out of the instructors teaching Calculus II are graduate students. The rest of the instructors are a combination of lecturers, post doctoral students and non-tenure track faculty. Only approximately 2 percent of the instructors are tenure track faculty. As for Calculus III, 99.17 percent of the laboratories

[1]The Math sequences offered are: the standard Math sequence, the applied honors Calculus sequence, the honors Calculus sequence and the honors seminar Math sequence.

are taught by TAs while the lectures are taught by other types of instructors (professors, lecturers, etc.). All of the three courses considered have uniform exam dates. The exams are not multiple choice, but the TAs grade the exams together using the same answer key.

Table A.5: Mathematics Courses Considered

Course	Components
Calculus I	LEC (taught by TA)
Calculus II	LEC (taught by TA)
Calculus III	LEC + LAB

A.2.2 Physics

The Physics Department offers three introductory course sequences. The Fundamental Concepts of Physics Sequence is comprised of the following courses: General Physics I, Elementary Lab I, General Physics II, Elementary Lab II, Waves Heat Light, Waves, Heat and Light Lab(see Table A.6). General Physics I covers classical mechanics, while General Physics II covers electricity, magnetism, optics, and introduces concepts in modern physics. This sequence is designed for prospective physical science and engineering undergraduate students.[2] All the Physics laboratories have grades separate from the courses they pertain to. The exams for introductory classes take place at the same time. Most of the introductory courses have three midterms and a final exam.

General Physics I and II both consist of a lecture and a discussion session. Since only 4.46 of the discussion sessions in General Physics I are taught by TAs and none of the ones for General Physics II is taught by TAs (they are taught by a combination of professors (full, assistant or associate) and lecturers), I do not consider these two course for my analysis. Elementary Lab I is taught by 84.97 graduate students and it is a two-hour weekly laboratory designed to accompany General Physics I. Elementary Lab II contains a two-hour weekly laboratory that is taken at the same time with General Physics II. 84.69 of the instructors for the Elementary Lab II are TAs. The grade of this course is based on class performance and laboratory reports submitted each lab session (10 lab experiments in total). The lab courses have multiple choice quizzes graded by each TA. In addition to these quizzes, each section also has laboratory worksheets that are graded on an answer key made by the TAs.

[2]The university also offers a sequence for prospective life sciences students and one for honors students.

Table A.6: Physics Courses Considered

Course	Components
Fundamental Concepts of Physics Sequence	
General Physics I	LEC+ DISC
Elementary Lab I	LAB
General Physics II	LEC+ DISC
Elementary Lab II	LAB

A.2.3 Chemistry

The general sequence (see Table A.7) for undergraduate students interested in the sciences, engineering or medicine starts with either General Chemistry or Structure and Reactivity I, depending on how strong their Chemistry background is[3]. General Chemistry has a discussion and a lecture, where 97.68 percent of the discussions are taught by TAs. The General Chemistry Lab I consists of a discussion session and a laboratory, both taught mostly by TAs (95.82 percent and 95.26 respectively of TA-led sections). Structure and Reactivity I contains a discussion session (96.13 percent of discussions are taught by TAs) and a lecture. The course Investigations in Chemistry is made up of a laboratory and a lecture. Out of all the laboratory sessions, 78.09 percent are taught by TAs. The Synthesis and Characterization of Organic Compounds is composed of a lecture and a laboratory (85 percent taught by TAs). Structure and Reactivity II contains a lecture, laboratory and discussion session. The laboratory is taught in proportion of 84.35 by TAs and all discussion sessions are taught by TAs. There are no exams for the lab courses and the laboratory reports are graded by each TA using an answer key. The exams for the General Chemistry are multiple choice and scantron-graded, while the exams for Structure and Reactivity II are not multiple choice and grading is done together by all the TAs teaching the course, using a grading system set by the professor teaching the lecture.

Table A.7: Chemistry Courses Considered

Course	Components
General Chemistry	LEC+DISC
General Chemistry Lab I	LAB
Structure and Reactivity I	LEC+DISC
Investigations in Chemistry: Laboratory	LEC+LAB
Structure and Reactivity II	LEC
The Synthesis and Characterization of Organic Compounds	LAB+LEC

[3]Students who took Chemistry AP credits in high school are advised to start with Structure and Reactivity I.

A.2.4 Biology

Undergraduate students interested in majoring in biological sciences take the Introductory Biology Sequence (see Table A.8). The first Biology course, Ecology/Evolution and Molecular, contains a lecture and a discussion session and 78.88 percent of the discussion sessions are led by TAs. The second Biology course, Introductory Biology - Molecular, Cellular, and Developmental, contains a lecture and a laboratory. The majority of the discussion sessions are taught by TAs (71.96 percent). These two courses are supplemented by an Introductory Biology Lab, with is taught by TAs almost entirely (94.97 of them are TA-led). Both of the Introductory Biology courses (Ecology and Evolution/ Molecular, Cellular, and Developmental) have multiple choice, scantron-grade exams (with the possibility of some short answers as well). The Introductory Biology laboratory contains two quizzes graded by each TA using an answer key provided by the lecture instructor. The older Biology introductory course contains both a discussion and a laboratory, both taught by the same TA, with 93.37 of labs/discussion sessions led by TAs.

Table A.8: Biology Courses Considered

Course	Components
Introductory Biology Sequence	
Introductory Biology- Ecology and Evolution	LEC+DISC
Introductory Biology - Molecular, Cellular, and Developmental	LEC+DISC
Introductory Molecular Bio-Engineering	LEC
Introductory Biology Lab	LAB
Older courses	
Introductory Biology	LEC+DISC+LAB
Honors Introductory Biology	LEC+DISC
Introductory Microbiology	LEC+LAB

A.2.5 Engineering

All undergraduate students planning to major in Engineering are required to take a multitude of courses in different fields, including Mathematics, Physics, Chemistry and Engineering. Two of the required courses in any first year Engineering program are Introduction to Engineering and Introduction to Computing and Programming (see Table A.9). While Introduction to Engineering contains both a discussion session and a laboratory, only $25,58$ percent of the labs and only 4.87 percent of the discussions sessions are taught by TAs. Therefore, I disregard this course and only consider Introduction to Computing and Programming, where TAs led 90.36 of the laboratories. Another reason for not including

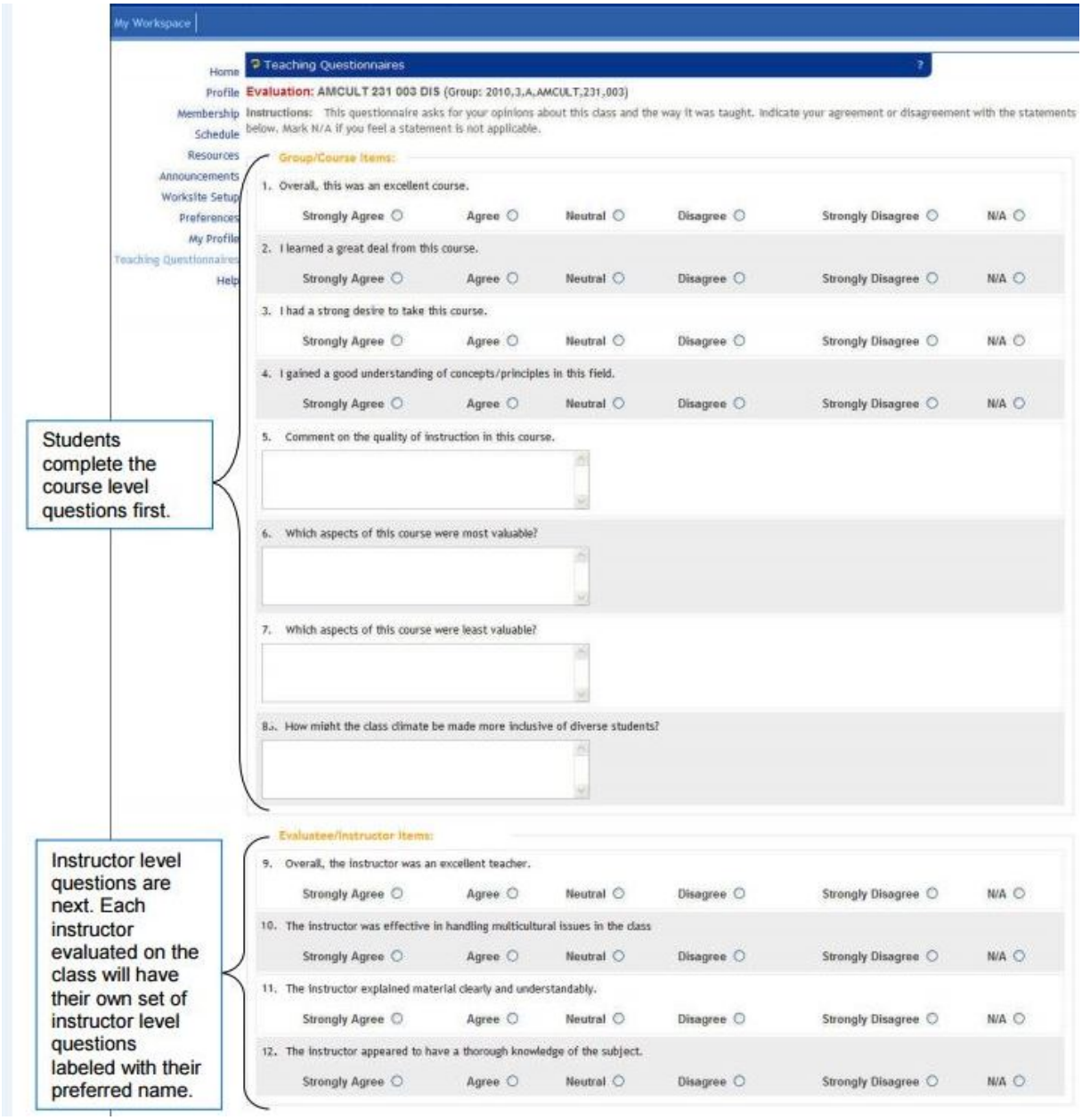
the Introduction to Engineering course is that the course does not have exams, but rather team projects, making the course too different than all the other courses considered in my analysis. The Introduction to Computing and Programming course has exams that are a combination of multiple choice questions and shorts answers and are graded by the TAs and the professors.

Table A.9: Engineering Courses Considered

Course	Components
Introduction to Engineering	LEC+DISC+LAB
Introduction to Computing and Programming	LEC+LAB

A.3 Student evaluation of teaching questions

Figure A.1: Student evaluation of teaching questionnaire



A.4 Computational example for the calculation of the median evaluation score

This section illustrates the calculation of the median score for each evaluation question based on the teaching evaluations filled out by the undergraduate students in each course.

Table A.10: Example of student evaluation scores

Score	1	2	3	4	5
f	3	8	2	5	1
cf	3	11	13	18	19

Table A.10 shows the student evaluation answers for a hypothetical question, the frequency and cumulative frequency of these answers. The median is defined as the point where or below where exactly 50 percent of the cases fall (Hays, 1973). This implies that the frequency at the median should be exactly half of the total number of observations. Based on this, the median would divide the distribution into halves, with 19/2 scores above and 19/2 scores below the median. The scores don't quite divide themselves into two groups, and as seen above the median would fall somewhere in the interval containing 2. The upper and lower limits of this interval are 1.5 and 2.5, respectively. The median calculation is determined by interpolation by using the following formula:

$$m = L + c\frac{\frac{N}{2} - F_m}{b} \tag{A.1}$$

In the above formula, m =median, L =lower limit of the interval containing the median, c =the width of the interval containing the median=upper real limit−lower real limit, N =total number of responses, F = cumulative frequency b =number of observations within the interval containing the median. This implies:

$$\text{Median} = 1.5 + 1 * \frac{\frac{19}{2} - 3}{8} = 2.31 \tag{A.2}$$

Table A.11: Sensitivity to the inclusion of different controls

Median evaluation scores			
Discussion sessions			
Foreign TA from non-English speaking country	-0.36**	-0.36**	-0.49***
	(0.13)	(0.12)	(0.13)
Foreign TA from English speaking country	-0.35	-0.31	-0.36
	(0.19)	(0.21)	(0.21)
Laboratories			
Foreign TA from non-English speaking country	-0.24*	-0.21*	-0.33**
	(0.10)	(0.10)	(0.10)
Foreign TA from English speaking country	-0.05	-0.06	-0.09
	(0.21)	(0.21)	(0.21)
Full courses			
Foreign TA from non-English speaking country	-0.52***	-0.58***	-0.48**
	(0.16)	(0.17)	(0.14)
Foreign TA from English speaking country	-0.27	-0.27	-0.30*
	(0.15)	(0.17)	(0.15)
Undergraduate student controls	Yes	No	Yes
TA controls	Yes	Yes	No
Grades			
Discussion sessions			
Foreign TA from non-English speaking country	-0.04	-0.01	0.01
	(0.02)	(0.03)	(0.02)
Foreign TA from English speaking country	-0.05	-0.12	-0.00
	(0.03)	(0.06)	(0.03)
Laboratories			
Foreign TA from non-English speaking country	-0.03	-0.04	0.01
	(0.02)	(0.02)	(0.01)
Foreign TA from English speaking country	-0.04	-0.05	0.00
	(0.03)	(0.03)	(0.03)
Full courses			
Foreign TA from non-English speaking country	-0.03	-0.06	-0.03
	(0.03)	(0.04)	(0.03)
Foreign TA from non-English speaking country	-0.04	-0.08	-0.04
	(0.05)	(0.05)	(0.05)

Notes: All specifications control for TA gender, race, age, times taught before, section time, and day of section. Course and term fixed effects are included. The median evaluation scores regressions also control for average undergraduate student characteristics and have the the standard errors are clustered by TA. The grade regressions control for undergraduate student characteristics and have the the standard errors two-way clustered (undergraduate student and TA level). Standard errors in parentheses.

*** Statistical significance at the 1 percent level.

** Statistical significance at the 5 percent level.

* Statistical significance at the 10 percent level.

BIBLIOGRAPHY

Aakvik, A. 2001. "Bounding a matching estimator: the case of a Norwegian training program." *Oxford bulletin of economics and statistics*, 63(1): 115–143.

Abadie, A., and G.W. Imbens. 2012. "A martingale representation for matching estimators." *Journal of the American Statistical Association*, 107(498): 833–843.

Altonji, J.G., T.E. Elder, and C.R. Taber. 2005. "An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling." *Journal of Human Resources*, 40(4): 791–821.

Andersen, K., and E.D. Miller. 1997. "Gender and student evaluations of teaching." *Political Science and Politics*, 30: 216–218.

Andrade, M.S. 2006. "International students in English-speaking universities: Adjustment factors." *Journal of Research in International education*, 5(2): 131–154.

Arcidiacono, P. 2004. "Ability sorting and the returns to college major." *Journal of Econometrics*, 121(1-2): 343–375.

Arcidiacono, P., E.M. Aucejo, and K. Spenner. 2012. "What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice." *IZA Journal of Labor Economics*, 1: 1–24.

Arcidiacono, P., E.M. Aucejo, and V.J. Hotz. 2016. "University differences in the graduation of minorities in STEM fields: Evidence from California." *American Economic Review*, 106(3): 525–62.

Aronson, J., and M.S. McGlone. 2009. "The handbook of prejudice, stereotyping, and discrimination." , ed. T.D. Nelson, Chapter Stereotype and social identity threat, 153–178. New York: Psychology Press.

Augurzky, B., and C.M. Schmidt. 2001. "The propensity score: A means to an end." IZA Discussion Paper No. 271.

Augustine, N.R. 2007. "Rising above the gathering storm: Energizing and employing America for a brighter economic future." *Washington DC: the National Academies Press*, 19: 2007.

Bandura, A. 1982. "Self-efficacy mechanism in human agency." *American Psychologist*, 37(2): 122–147.

Barlow, A.E.L., and M. Villarejo. 2004. "Making a difference for minorities: Evaluation of an educational enrichment program." *Journal of Research in Science Teaching*, 41(9): 861–881.

Basow, S.A. 1995. "Student evaluations of college professors: When gender matters." *Journal of Educational Psychology*, 87(4): 656–665.

Basow, S.A., and N.T. Silberg. 1987. "Student evaluations of college professors: Are female and male professors rated differently?" *Journal of Educational Psychology*, 79: 308–314.

Basow, S.A., J.E. Phelan, and L. Capotosto. 2006. "Gender patterns in college students' choices of their best and worst professors." *Psychology of Women Quarterly*, 30(1): 25–35.

Bauer, K.W., and J.S. Bennett. 2003. "Alumni perceptions used to assess undergraduate research experience." *The Journal of Higher Education*, 74: 210–230.

Becker, S.O., and M. Caliendo. 2007. "Sensitivity analysis for average treatment effects." *The Stata Journal*, 7(1): 71–83.

Beede, D.N., T.A. Julian, D. Langdon, G. McKittrick, B. Khan, and M.E. Doms. 2011. "Women in STEM: A gender gap to innovation." *U.S. Department of Commerce*. Economics and Statistics Administration Issue Brief# 04-11.

Bettinger, E. 2010. "To be or not to be: Major choices in budding scientists." In *American universities in a global market*. 69–98. University of Chicago Press.

Bettinger, E.P., and B.T. Long. 2005. "Do faculty serve as role models? The impact of instructor gender on female students." *American Economic Review*, 95(2): 152–157.

Bettinger, E.P., and B.T. Long. 2010. "Does cheaper mean better? The impact of using adjunct instructors on student outcomes." *The Review of Economics and Statistics*, 92(3): 598–613.

Bettinger, E.P., B.T. Long, P. Oreopoulos, and L. Sanbonmatsu. 2009. "The role of simplification and information in college decisions: Results from the H&R Block FAFSA experiment." National Bureau of Economic Research.

Bianchini, S., F. Lissoni, and Pezzoni. M. 2013. "Instructor characteristics and students' evaluations of teaching effectiveness." *European Journal of Engineering Education*, 38(1): 38–57.

Biernat, M., M. Manis, and T.E. Nelson. 1991. "Stereotypes and standards of judgment." *Journal of Personality and Social Psychology*, 60(4): 485–499.

Black, D., K. Daniel, and J. Smith. 2005. "College quality and wages in the United States." *German Economic Review*, 6(3): 415–443.

Blau, F.D., and L.M. Kahn. 2016. "Collective bargaining, relative wages, and employment." *The Changing Role of Unions: New Forms of Representation: New Forms of Representation*.

Blickenstaff, J.C. 2005. "Women and science careers: Leaky pipeline or gender filter?" *Gender and Education*, 17(4): 369–386.

Boring, A. 2017. "Gender biases in student evaluations of teaching." *Journal of Public Economics*, 145: 27–41.

Boring, A., K. Ottoboni, and P. Stark. 2016. "Student evaluations of teaching (mostly) do not measure teaching effectiveness."

Borjas, G.J. 2000. "Foreign-born teaching assistants and the academic performance of undergraduates." *The American Economic Review*, 90(2): 355–359.

Boudarbat, B., and C. Montmarquette. 2009. "Choice of fields of study of Canadian university graduates: the role of gender and their parents' education." *Education Economics*, 17(2): 185–213.

Bound, J., S. Turner, and P. Walsh. 2009. "Internationalization of US doctorate education." In *Science and Engineering Careers in the United States: An Analysis of Markets and Employment*. 59–97. University of Chicago Press.

Bowman, N.A. 2010. "Can 1st-year college students accurately report their learning and development?" *American Educational Research Journal*, 47(2): 466–496.

Braga, M., M. Paccagnella, and M. Pellizzari. 2014. "Evaluating students' evaluations of professors." *Economics of Education Review*, 41: 71–88.

Braga, M., M. Paccagnella, and M. Pellizzari. 2016. "The impact of college teaching on students' academic and labor market outcomes." *Journal of Labor Economics*, 34(3): 781–822.

Brodaty, T., and M. Gurgand. 2016. "Good peers or good teachers? Evidence from a French University." *Economics of Education Review*, 54: 62–78.

Brown, C., and M. Corcoran. 1997. "Sex-based differences in school content and the male-female wage gap." *Journal of Labor Economics*, 15(3): 431–465.

Bryson, A., R. Dorsett, and S. Purdon. 2002. "The use of propensity score matching in the evaluation of labour market policies." Working Paper No. 4, Department for Work and Pensions.

Bureau of Labor Statistics, U.S. Department of Labor. 2016. "Occupational Employment Statistics."

Busso, M., J. DiNardo, and J. McCrary. 2014. "New evidence on the finite sample properties of propensity score reweighting and matching estimators." *Review of Economics and Statistics*, 96(5): 885–897.

Butler, D.M., and R. Christensen. 2003. "Mixing and matching: The effect on student performance of teaching assistants of the same gender." *Political Science and Politics*, 36(4): 781–786.

Cameron, A.C., J.B. Gelbach, and D.L. Miller. 2011. "Robust inference with multiway clustering." *Journal of Business and Economic Statistics*, 29(2): 238–249.

Canes, B.J., and H.S. Rosen. 1995. "Following in her footsteps? Women's choices of college majors and faculty gender composition." *Industrial and Labor Relations Review*, 48(3): 486–504.

Carlone, H.B., and A. Johnson. 2007. "Understanding the science experiences of successful women of color: Science identity as an analytic lens." *Journal of Research in Science Teaching*, 44(8): 1187–1218.

Carrell, S.E., and J.E. West. 2010. "Does professor quality matter? Evidence from random assignment of students to professors." *Journal of Political Economy*, 118(3): 409–432.

Carrell, S.E., M.E. Page, and J.E. West. 2010. "Sex and science: How professor gender perpetuates the gender gap." *The Quarterly Journal of Economics*, 125(3): 1101–1144.

Cech, E., B. Rubineau, S. Silbey, and C. Seron. 2011. "Professional role confidence and gendered persistence in engineering." *American Sociological Review*, 76(5): 641–666.

Ceci, S.J., and W.M. Williams. 2009. *The mathematics of sex: How biology and society conspire to limit talented women and girls.* New York: Oxford University Press.

Ceci, S.J., and W.M. Williams. 2010. "Sex differences in math-intensive fields." *Current Directions in Psychological Science*, 19: 275–279.

Centra, J.A., and N.B. Gaubatz. 2000. "Is There Gender Bias in Student Evaluations of Teaching?" *The Journal of Higher Education*, 70: 17–33.

Chang, M.J., J. Sharkness, S. Hurtado, and C.B. Newman. 2014. "What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups." *Journal of Research in Science Teaching*, 51(5): 555–580.

Chen, X. 2013. "STEM attrition: College students' paths into and out of STEM Fields (NCES 2014-001)." *Washington, Dc: US. Department of Education and National Center for Education Statistics*.

Chetty, R., J.N. Friedman, and J.E. Rockoff. 2014*a*. "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *The American Economic Review*, 104(9): 2593–2632.

Chetty, R., J.N. Friedman, and J.E. Rockoff. 2014*b*. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *The American Economic Review*, 104(9): 2633–2679.

Cole, D., and A. Espinoza. 2008. "Examining the academic success of Latino students in science technology engineering and mathematics (STEM) majors." *Journal of College Student Development*, 49(4): 285–300.

Corcoran, S.P., J.L. Jennings, and A.A. Beveridge. 2011. "Teacher effectiveness on high- and low-stakes tests." *Society for Research on Educational Effectiveness*.

Correia, S. 2016. "A feasible estimator for linear models with multi-way fixed effects."

Correll, S.J. 2001. "Gender and the career choice process: The role of biased self-assessments." *American Journal of Sociology*, 106(6): 1691–1730.

Cramer, K.M., and L.R. Alexitch. 2000. "Student evaluations of college professors: identifying sources of bias." *Canadian Journal of Higher Education*, 30(2): 143–164.

Crisp, G., A. Nora, and A. Taggart. 2009. "Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution." *American Educational Research Journal*, 46(4): 924–942.

Dalmia, S., D.C. Giedeman, H. A. Klein, and N.M. Levenburg. 2005. "Women in academia: An analysis of their expectations, performance and pay." *Forum on Public Policy*, 1: 160–177.

Deming, D., and S. Dynarski. 2009. "Into college, out of poverty? Policies to increase the postsecondary attainment of the poor." National Bureau of Economic Research.

De Vlieger, P., B. Jacob, and K. Stange. 2017. "Measuring instructor effectiveness in higher education." In *Productivity in Higher Education*. , ed. C.M. Hoxby and K. Stange. University of Chicago Press.

DiPrete, T.A., and C. Buchmann. 2013. *The rise of women: The growing gender gap in education and what it means for American schools.* Russell Sage Foundation, New York.

DiPrete, T.A., and M. Gangl. 2004. "Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments." *Sociological Methodology*, 34(1): 271–310.

Dunning, D., K. Johnson, J. Ehrlinger, and J. Kruger. 2003. "Why people fail to recognize their own incompetence." *Current directions in Psychological Science*, 12(3): 83–87.

Dynan, K.E., and C.E. Rouse. 1997. "The underrepresentation of women in economics: A study of undergraduate economics students." *The Journal of Economic Education*, 28(4): 350–368.

Eddy, S.L., S.E. Brownell, and M.P. Wenderoth. 2014. "Gender gaps in achievement and participation in multiple introductory biology classrooms." *CBE-Life Sciences Education*, 13(3): 478–492.

Ehrenberg, R.G. 2005. "Graduate education, innovation and federal responsibility." *Communicator*, 38: 1–8.

Ehrenberg, R.G. 2010. "Analyzing the factors that influence persistence rates in STEM field, majors: Introduction to the symposium." *Economics of Education Review*, 29(6): 888–891.

Ehrenberg, R.G., D.D. Goldhaber, and D.J. Brewer. 1995. "Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988." *Industrial and Labor Relations Review*, 48(3): 547–561.

Ellington, R. 2006. "Having their say: Eight high-achieving African-American undergraduate mathematics majors discuss their success and persistence in mathematics." PhD diss. University of Maryland, College Park.

England, P., and S. Li. 2006. "Desegregation stalled the changing gender composition of college majors, 1971–2002." *Gender & Society*, 20(5): 657–677.

England, P., P. Allison, S. Li, N. Mark, J. Thompson, M.J. Budig, and H. Sun. 2007. "Why are some academic fields tipping toward female? The sex composition of US fields of doctoral degree receipt, 1971-2002." *Sociology of Education*, 80(1): 23–42.

Fairlie, R.W., F. Hoffmann, and P. Oreopoulos. 2014. "A community college instructor like me: Race and ethnicity interactions in the classroom." *The American Economic Review*, 104(8): 2567–2591.

Feldman, K.A. 1993. "College students' views of male and female college teachers: Part II. Evidence from students' evaluations of their classroom teachers." *Research in Higher Education*, 34: 151–211.

Feldon, D.F., M.A. Maher, M. Hurst, and B. Timmerman. 2015. "Faculty mentors, graduate students, and performance-based assessments of students research skill development." *American Educational Research Journal*, 52(2): 334–370.

Figlio, D.N., M.O. Schapiro, and K.B. Soter. 2015. "Are tenure track professors better teachers?" *Review of Economics and Statistics*, 97(4): 715–724.

Fleisher, B., M. Hashimoto, and B. Weinberg. 2002. "Foreign GTAs can be effective teachers of Economics." *The Journal of Economic Education*, 33(4): 299–325.

Foschi, M. 2000. "Double standards for competence: Theory and research." *Annual Review of Sociology*, 26: 21–42.

Frölich, M. 2004. "Finite sample properties of propensity-score matching and weighting estimators." *Review of Economics*, 86: 77–90.

Gaulé, P., and M. Piacentini. 2013. "Chinese graduate students and US scientific productivity." *Review of Economics and Statistics*, 95(2): 698–701.

Gayles, J.G., and F. Ampaw. 2014. "The impact of college experiences on degree completion in STEM fields at four-year institutions: Does gender matter?" *The Journal of Higher Education*, 85(4): 439–468.

Goldin, C. 2006. "The quiet revolution that transformed women's employment, education, and family." *American Economic Review*, 96(2): 1–21.

Goldin, C. 2013. "Notes on women and the undergraduate economics major." *CSWEP Newsletter*, 2013(15): 4–6.

Goldin, C., and L.F. Katz. 2008. "Transitions: Career and family life cycles of the educational elite." *The American Economic Review*, 98(2): 363–369.

Gonzalez, H.B., and J.J. Kuenzi. 2012. "Science, technology, engineering, and mathematics (STEM) education: a primer." *Congressional Research Service*. Washington, D.C.

Good, C., A. Rattan, and C.S. Dweck. 2012. "Why do women opt out? Sense of belonging and women's representation in mathematics." *Journal of Personality and Social Psychology*, 102(4): 700.

Goyette, K., and Y. Xie. 1999. "Educational expectations of Asian American youths: Determinants and ethnic differences." *Sociology of Education*, 22–36.

Graham, M.J., A. Byars-Winston, A.B. Hunter, and J. Handelsman. 2013. "Science education: Increasing persistence of college students in STEM." *Science*, 341: 1455–1456.

Grandy, J. 1998. "Persistence in science of high-ability minority students: Results of a longitudinal study." *The Journal of Higher Education*, 69(6): 589–620.

Gregerman, S.R., J.S. Lerner, W. von Hippel, J. Jonides, and B.A. Nagda. 1998. "Undergraduate student-faculty research partnerships affect student retention." *The Review of Higher Education*, 22(1): 55–72.

Griffith, A.L. 2010. "Persistence of women and minorities in STEM field majors: Is it the school that matters?" *Economics of Education Review*, 29(6): 911–922.

Griffith, A.L. 2014. "Faculty gender in the college classroom: Does it matter for achievement and major choice?" *Southern Economic Journal*, 81(1): 211–231.

Grunspan, D.Z., S.L. Eddy, S.E. Brownell, B.L. Wiggins, A.J. Crowe, and S.M. Goodreau. 2016. "Males under-estimate academic performance of their female peers in undergraduate biology classrooms." *PloS one*, 11(2): e0148405.

Guiso, L., F. Monte, P. Sapienza, and L. Zingales. 2008. "Culture, gender, and math." *Science*, 320(5880): 1164–1165.

Hanushek, E. 1971. "Teacher characteristics and gains in student achievement: Estimation using micro data." *The American Economic Review*, 61(2): 280–288.

Hathaway, R.S., B.A. Nagda, and S.R. Gregerman. 2002. "The relationship of undergraduate research participation to graduate and professional education pursuit: An empirical study." *Journal of College Student Development*, 43(5): 614–631.

Hays, W.L. 1973. *Statistics for the Social Sciences.* . 2nd ed., New York: Holt, Rinehart and Winston.

Heckman, J., and J. Smith. 2004. "The determinants of participation in a social program: Evidence from a prototypical job training program." *Journal of Labor Economics*, 22(4): 333–372.

Heckman, J., and R. Robb. 1985. "Longitudinal analysis of labor market data." , ed. James H. and Burton S., Chapter Alternative Methods for Evaluating the Impact of Interventions, 156–246. New York: Cambridge University Press.

Heckman, J., H. Ichimura, and P. Todd. 1997. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme." *The Review of Economic Studies*, 64(4): 605–654.

Hellerstein, J.K., and M.S. Morrill. 2011. "Dads and daughters the changing impact of fathers on women's occupational choices." *Journal of Human Resources*, 46(2): 333–372.

Hernandez, P.R., A. Woodcock, M. Estrada, and P.W. Schultz. 2018. "Undergraduate research experiences broaden diversity in the scientific workforce." *BioScience*, 68(3): 204–211.

Herrera, F.A., and S. Hurtado. 2011. "Maintaining initial interests: Developing science, technology, engineering, and mathematics (STEM) career aspirations among underrepresented racial minority students."

Hill, C., C. Corbett, and A. St Rose. 2010. *Why so few? Women in Science, Technology, Engineering, and Mathematics.* American Association of University Women, Washington, DC.

Hoffmann, F., and P. Oreopoulos. 2009. "A professor like me: The influence of instructor gender on college achievement." *Journal of Human Resources*, 44(2): 479–494.

Hossler, D., M. Ziskin, J.P.K. Gross, S. Kim, and O. Cekic. 2009. "Student aid and its role in encouraging persistence." In *Higher education: Handbook of Theory and Research*. 389–425.

Huber, M., M. Lechner, and C. Wunsch. 2013. "The performance of estimators based on the propensity score." *Journal of Econometrics*, 175: 1–21.

Hunter, A.B., S.L. Laursen, and E. Seymour. 2007. "Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development." *Science Education*, 91(1): 36–74.

Hunt, J. 2016. "Why do women leave science and engineering?" *ILR Review*, 69(1): 199–226.

Hurtado, S., N.L. Cabrera, M.H. Lin, L. Arellano, and L.L. Espinosa. 2009. "Diversifying science: Underrepresented student experiences in structured research programs." *Research in Higher Education*, 50(2): 189–214.

Hu, S., G.D. Kuh, and J.G. Gayles. 2007. "Engaging undergraduate students in research activities: Are research universities doing a better job?" *Innovative Higher Education*, 32(3): 167–177.

Jackson, C.K. 2013. "Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina." NBER Working Paper No. 18624.

Jacobs, L.C., and C.B. Friedman. 1988. "Student achievement under foreign teaching associates compared with native teaching associates." *The Journal of Higher Education*, 59(5): 551–563.

Johnson, A.C. 2007. "Unintended consequences: How science professors discourage women of color." *Science Education*, 91(5): 805–821.

Johnson, H.J., L. Barnard-Brak, T.F. Saxon, and M.K. Johnson. 2012. "An experimental study of the effects of stereotype threat and stereotype lift on men and women's performance in mathematics." *The Journal of Experimental Education*, 80(2): 137–149.

Kane, T.J., and D.O. Staiger. 2008. "Estimating teacher impacts on student achievement: An experimental evaluation." National Bureau of Economic Research. NBER Working Paper 1460.

Kao, G., and J.S. Thompson. 2003. "Racial and ethnic stratification in educational achievement and attainment." *Annual Review of Sociology*, 29: 417–442.

Kardash, C.M. 2000. "Evaluation of an undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors." *Journal of Educational Psychology*, 92(1): 191–201.

Kenney-Benson, G.A., E.M. Pomerantz, A.M. Ryan, and H. Patrick. 2006. "Sex differences in math performance: The role of children's approach to schoolwork." *Developmental Psychology*, 42(1): 11–26.

Kim, M.M., G. Rhoades, and D.B. Woodard Jr. 2003. "Sponsored research versus graduating students? Intervening variables and unanticipated findings in public research universities." *Research in Higher Education*, 44(1): 51–81.

Kinkead, J. 2003. "Learning through inquiry: An overview of undergraduate research." *New Directions for Teaching and Learning*, 93: 5–17.

Koedel, C., K. Mihaly, and J.E. Rockoff. 2015. "Value-added modeling: A review." *Economics of Education Review*, 47: 180–195.

Koester, B.P., G. Grom, and T.A. McKay. 2016. "Patterns of gendered performance difference in introductory STEM courses."

Kokkelenberg, E.C., and E. Sinha. 2010. "Who succeeds in STEM studies? An analysis of Binghamton University undergraduate students." *Economics of Education Review*, 29(6): 935–946.

Konstantopoulos, S., and V. Chung. 2011. "The persistence of teacher effects in elementary grades." *American Educational Research Journal*, 48(2): 361–386.

Krautmann, A.C., and W. Sander. 1999. "Grades and student evaluations of teachers." *Economics of Education Review*, 18(1): 59–63.

Linn, M.C., E. Palmer, A. Baranger, E. Gerard, and E. Stone. 2015. "Undergraduate research experiences: Impacts and opportunities." *Science*, 347(6222). art. 1261757.

Lopatto, D. 2004. "Survey of undergraduate research experiences (SURE): First findings." *Cell Biology Education*, 3(4): 270–277.

Lopatto, D. 2010. "Undergraduate research as a high-impact student experience." *Peer Review*, 12(2): 27–30.

Lusher, L., D. Campbell, and S. Carrell. 2015. "TAs like me: Racial interactions between graduate teaching assistants and undergraduates." National Bureau of Economic Research.

MacLachlan, Anne J. 2006. "Developing graduate students of color for the professoriate in Science, Technology, Engineering, and Mathematics (STEM)." *Center for Studies in Higher Education*. Research & Occasional Paper Series: CSHE. 6.06, University of California, Berkeley.

MacNell, L., A. Driscoll, and A.N. Hunt. 2015. "What's in a name: exposing gender bias in student ratings of teaching." *Innovative Higher Education*, 40(4): 291–303.

Maltese, A.V., and R.H. Tai. 2010. "Eyeballs in the fridge: Sources of early interest in science." *International Journal of Science Education*, 32(5): 669–685.

Maltese, A.V., and R.H. Tai. 2011. "Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among US students." *Science Education*, 95(5): 877–907.

Mann, A., and T.A. DiPrete. 2013. "Trends in gender segregation in the choice of science and engineering majors." *Social science research*, 42(6): 1519–1541.

Maple, S.A., and F.K. Stage. 1991. "Influences on the choice of math/science major by gender and ethnicity." *American Educational Research Journal*, 28(1): 37–60.

Mau, W.C. 2003. "Factors that influence persistence in science and engineering career aspirations." *The Career Development Quarterly*, 51(3): 234–243.

Meinholdt, C., and S.L. Murray. 1999. "Why aren't there more women engineers?" *Journal of Women and Minorities in Science and Engineering*, 5: 239–263.

Mengel, F., J. Sauermann, and U. Zölitz. 2017. "Gender bias in teaching evaluations." IZA Discussion Paper No. 11000.

Mervis, J. 2006. "Biomedical training programs: NIH told to get serious about giving minorities a hand." *Science*, 311: 328–329.

Micari, M., P. Pazos, and M.J.Z. Hartmann. 2007. "A matter of confidence: Gender differences in attitudes toward engaging in lab and course work in undergraduate engineering." *Journal of Women and Minorities in Science and Engineering*, 13(3).

Miller, J., and M. Chamberlin. 2000. "Women are teachers, men are professors: A study of student perceptions." *Teaching Sociology*, 28(4.): 283–298.

Miller, J.D., and L.G. Kimmel. 2012. "Pathways to a STEMM Profession." *Peabody Journal of Education*, 87(1): 26–45.

Murnane, R.J. 1975. *The impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger Publishing.

Murphy, M.C., C.M. Steele, and J.J. Gross. 2007. "Signaling threat: How situational cues affect women in math, science, and engineering settings." *Psychological Science*, 18(10): 879–885.

Murray, H.G. 2005. "Student evaluation of teaching: Has it made a difference?"

National Science Board. 2016. *Science and engineering indicators 2016 (NSB-2016-1)*. Arlington, VA: National Science Foundation.

National Science Foundation, National Center for Science, and Engineering Statistics. 2017. "Women, minorities, and persons with disabilities in science and engineering: 2017."

Nguyen, H.H.D., and A.M. Ryan. 2008. "Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence." *Journal of applied psychology*, 93(6): 1314–1334.

Nollenberger, N., N. Rodríguez-Planas, and A. Sevilla. 2016. "The math gender gap: The role of culture." *American Economic Review*, 106(5): 257–61.

Norris, T. 1991. "Nonnative English-speaking teaching assistants and student performance." *Research in Higher Education*, 32(4): 433–448.

NSB. 2014. *Science & Engineering Indicators.* Arlington, VA: Natl. Cent. Sci. Eng. Stat.

Nye, B., S. Konstantopoulos, and L.V. Hedges. 2004. "How large are teacher effects?" *Educational evaluation and policy analysis*, 26(3): 237–257.

Olson, S., and D.G. Riordan. 2012. "Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President." *Executive Office of the President*.

Ost, B. 2010. "The role of peers and grades in determining major persistence in the sciences." *Economics of Education Review*, 29(6): 923–934.

Papay, J.P. 2011. "Different tests, different answers: The stability of teacher value-added estimates across outcome measures." *American Educational Research Journal*, 48(1): 163–193.

Pascarella, E.T., and P.T. Terenzini. 1979. "Interaction effects in Spady and Tinto's conceptual models of college attrition." *Sociology of Education*, 52: 197–210.

Pascarella, E.T., and P.T. Terenzini. 2005. *How college affects students: A third decade of research (Vol. 2)*. San Francisco: Jossey-Bass.

Pender, M., D.E. Marcotte, M.R.S. Domingo, and K.I. Maton. 2010. "The STEM pipeline: The role of summer research experience in minority students' Ph.D. aspirations." *Education Policy Analysis Archives*, 18(30): 1–19.

Perna, L., V. Lundy-Wagner, N.D. Drezner, M. Gasman, S. Yoon, E. Bose, and S. Gary. 2009. "The contribution of HBCUs to the preparation of African American women for STEM careers: A case study." *Research in Higher Education*, 50(1): 1–23.

Plakans, B.S. 1997. "Undergraduates' experiences with and attitudes toward international teaching assistants." *TESOL quarterly*, 31(1): 95–119.

Pope, D.G., and J.R. Sydnor. 2010. "Geographic variation in the gender differences in test scores." *Journal of Economic Perspectives*, 24(2): 95–108.

Price, J. 2010. "The effect of instructor race and gender on student persistence in STEM fields." *Economics of Education Review*, 29(6): 901–910.

Prieto, L., and E. Altmaier. 1994. "The relationship of prior training and previous teaching experience to self-efficacy among graduate teaching assistants." *Research in Higher Education*, 35(4): 481–497.

Qian, Y., B. Zafar, and H. Xie. 2009. "Do female faculty influence female students' choice of college major, and why?" Northwestern University Working Paper.

Rask, K. 2010. "Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences." *Economics of Education Review*, 29(6): 892–900.

Rask, K., and J. Tiefenthaler. 2008. "The role of grade sensitivity in explaining the gender imbalance in undergraduate economics." *Economics of Education Review*, 27(6): 676–687.

Rask, K.N., and E.M. Bailey. 2002. "Are faculty role models? Evidence from major choice in an undergraduate institution." *The Journal of Economic Education*, 33(2): 99–124.

Rivkin, S.G., E.A. Hanushek, and J.F. Kain. 2005. "Teachers, schools, and academic achievement." *Econometrica*, 73(2): 417–458.

Robst, J., J. Keil, and D. Russo. 1998. "The effect of gender composition of faculty on student retention." *Economics of Education Review*, 17(4): 429–439.

Rockoff, J.E. 2004. "The impact of individual teachers on student achievement: Evidence from panel data." *The American Economic Review*, 94(2): 247–252.

Rosen, A.S. 2017. "Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data." *Assessment & Evaluation in Higher Education*, 1–14.

Rosenbaum, P., and D. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41–55.

Rosenbaum, P.R. 2002. "Observational studies." In *Observational Studies*. . second ed., 1–17. Springer.

Rosenbaum, P.R., and D.B. Rubin. 1985. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." *The American Statistician*, 39: 33–38.

Rothstein, D.S. 1995. "Do female faculty influence female students' educational and labor market attainments?" *Industrial and Labor Relations Review*, 48(3): 515–530.

Rothstein, J. 2010. "Teacher quality in educational production: Tracking, decay, and student achievement." *The Quarterly Journal of Economics*, 125(1): 175–214.

Russell, S.H., M.P. Hancock, and J. McCullough. 2007. "The pipeline: Benefits of undergraduate research experiences." *Science*, 316(5824): 548–549.

Ryan, C.L., and K. Bauman. 2016. "Educational attainment in the United States: 2015."

Sabatini, D.A. 1997. "Teaching and research synergism: The undergraduate research experience." *Journal of Professional Issues in Engineering Education and Practice*, 123(3): 98–102.

Sadler, T., S. Burgin, L. McKinney, and L. Ponjuan. 2010. "Learning science through research apprenticeships: A critical review of the literature." *Journal of Research in Science Teaching*, 47: 235–256.

Sax, L.J. 1994. "Predicting gender and major-field differences in mathematical self-concept during college." *Journal of Women and Minorities in Science and Engineering*, 1(4): 291–307.

Sax, L.J. 2008. *The gender gap in college: Maximizing the developmental potential of women and men.* Jossey-Bass.

Schneider, B., C.B. Swanson, and C. Riegle-Crumb. 1998. "Opportunities for learning: Course sequences and positional advantages." *Social Psychology of Education*, 2: 25–53.

Scott-Clayton, J. 2011. "The causal effect of federal work-study participation: Quasi-experimental evidence from West Virginia." *Educational Evaluation and Policy Analysis*, 33(4): 506–527.

Scott-Clayton, J., and V. Minaya. 2016. "Should student employment be subsidized? Conditional counterfactuals and the outcomes of work-study participation." *Economics of Education Review*, 52: 1–18.

Seymour, E., A.B. Hunter, S.L. Laursen, and T. DeAntoni. 2004. "Establishing the benefits of research experiences for undergraduates in the sciences: first findings from a three-year study." *Science Education*, 88(4): 493–534.

Seymour, E., and N.M. Hewitt. 1997. *Talking about Leaving: Why Undergraduates Leave the Sciences*. Westview Press Boulder, Colorado.

Shannon, D., D. Twale, and M. Moore. 1998. "TA Teaching Effectiveness: The Impact of Training and Teaching Experience." *The Journal of Higher Education*, 69(4): 440–466.

Sidanius, J., and M. Crane. 1989. "Job evaluation and gender: The case of university faculty." *Journal of Applied Social Psychology*, 19: 174–197.

Smith, J., and P. Todd. 2005. "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics*, 125(1–2): 305–353.

Snyder, T.D., and S.A. Dillow. 2015. "Digest of education statistics 2013 (NCES 2015-011)." *National Center for Education Statistics, Institute of Education Sciences, US Department of Education*. Washington, DC: US Government Printing Office.

Spencer, S.J., C.M. Steele, and D.M. Quinn. 1999. "Stereotype threat and women's math performance." *Journal of experimental social psychology*, 35(1): 4–28.

Sprague, J., and K. Massoni. 2005. "Student evaluations and gendered expectations: What we can't count can hurt us." *Sex Roles*, 53: 779–793.

Stark, P.B., and R. Freishtat. 2014. "An evaluation of course evaluations." *Science Direct*.

Steele, C.M., and J. Aronson. 1995. "Stereotype threat and the intellectual test performance of African Americans." *Journal of Personality and Social Psychology*, 69(5): 797–811.

Stinebrickner, R., and T.R. Stinebrickner. 2003. "Working during school and academic performance." *Journal of Labor Economics*, 21(2): 473–491.

Stinebrickner, T.R., and R. Stinebrickner. 2013. "A major in science? Initial beliefs and final outcomes for college major and dropout." *The Review of Economic Studies*, 81(1): 426–472.

Stuart, E.A. 2010. "Matching methods for causal inference: A review and a look forward." *Statistical Science*, 25(1): 1.

Svinicki, M., and W.J. McKeachie. 2010. *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers.* . 13 ed., Belmont, CA: Wadsworth.

Tai, R.T., C.Q. Liu, A. Maltese, and X.T. Fan. 2006. "Planning early for careers in science." *Science*, 312(5777): 1143–1144.

Terenzini, P.T., and E.T. Pascarella. 1977. "Voluntary freshman attrition and patterns of social and academic integration in a university: A test of a conceptual model." *Research in Higher Education*, 6(1): 25–43.

Tinto, V. 1993. *Leaving College: Rethinking the causes and cures of student attrition.* . 2nd ed., University of Chicago Press, Chicago, IL.

Trice, A.G. 2003. "Faculty perceptions of graduate international students: The benefits and challenges." *Journal of Studies in International Education*, 7(4): 379–403.

Turner, S.E., and W.G. Bowen. 1999. "Choice of major: The changing (unchanging) gender gap." *Industrial and Labor Relations Review*, 52(2): 289–313.

Ullah, H., and M.A. Wilson. 2007. "Students' academic success and its association to student involvement with learning and relationships with faculty and peers." *College Student Journal*, 141(4): 1192–1202.

Urry, M. 2015. "Science and gender: scientists must work harder on equality." *Nature News*, 528(7583): 471–473.

US Department of Health and Human Services. 2015. "National Institute of General Medical Sciences 5-Year Strategic Plan."

Villarejo, M., and A.E.L. Barlow. 2007. "Evolution and evaluation of a biology enrichment program for minorities." *Journal of Women and Minorities in Science and Engineering*, 13(2).

Vogt, C.M., D. Hocevar, and L.S. Hagedorn. 2007. "A social cognitive construct validation: Determining women's and men's success in engineering programs." *The Journal of Higher Education*, 78(3): 337–364.

Ward, C., J.S. Bennett, and K.W. Bauer. 2003. "Content analysis of undergraduate research student evaluations." Retrieved June 2015.

Watts, M., and G.J. Lynch. 1989. "The principles courses revisited." *The American Economic Review*, 79(2): 236–241.

Weber, S., M. Appel, and N. Kronberger. 2015. "Stereotype threat and the cognitive performance of adolescent immigrants: The role of cultural identity strength." *Contemporary Educational Psychology*, 42: 71–81.

Weinberg, B.A., M. Hashimoto, and B.M. Fleisher. 2009. "Evaluating teaching in higher education." *The Journal of Economic Education*, 40(3): 227–261.

West, S.G, N. Duan, W. Pequegnat, P. Gaist, D.C. Des Jarlais, D. Holtgrave, J. Szapocznik, M. Fishbein, B. Rapkin, and M. Clatts. 2008. "Alternatives to the randomized controlled trial." *American Journal of Public Health*, 98(8): 1359–1366.

Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. n.d.. "Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit."

Wooldridge, J.M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.

World Bank Group. 2014. *World Development Indicators 2014*. World Bank Publications.

Worthington, A.C. 2002. "The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education." *Assessment and Evaluation in Higher Education*, 27(1): 49–64.

Xie, Y., and A.A. Killewald. 2012. *Is American science in decline?* Cambridge: Harvard University Press.

Xie, Y., and K.A. Shauman. 2003. *Women in science: Career processes and outcomes*. Harvard University Press.

Xie, Y., M. Fang, and K. Shauman. 2015. "STEM education." *Annual Review of Sociology*, 41: 331–357.

Zafar, B. 2013. "College major choice and the gender gap." *Journal of Human Resources*, 48(3): 545–595.

Zong, J., and J. Batalova. 2016. "International students in the United States." Migration Policy Institute.

Zydney, A.L., J.S. Bennett, A. Shahid, and K.W. Bauer. 2002. "Impact of undergraduate research experience in engineering." *Journal of Engineering Education*, 91(2): 151–157.